Dissertation

submitted to the

Combined Faculty of Natural Sciences and Mathematics

of the Ruperto Carola University Heidelberg, Germany

for the degree of

Doctor of Natural Sciences

Presented by

M. Sc. Stefano Secchia

born in Varese, Italy

Oral examination: October 8th, 2021

Single-cell dissection of regulatory landscapes during embryogenesis

Referees: Dr. Justin Crocker

Prof. Dr. Karsten Rippe

To my parents

Summary

Gene regulatory networks control the precise spatial and temporal execution of developmental gene expression programs during embryogenesis. Enhancers are core components of these regulatory networks and they serve as integration platforms for the developmental signals delivered by transcriptions factors (TFs). Dissecting the function of gene regulatory networks and their underlying components is a major goal of modern biology. In this thesis, I leverage the recent advances in single-cell genomics methods to assess the impact of perturbations on developmental regulatory networks, using *Drosophila melanogaster* embryogenesis as a well-studied model system.

I started by optimizing a protocol to profile single-cell chromatin accessibility by sci-ATAC-seq in *Drosophila* embryos and applied it to profile a dense time-course of over 20,000 cells during mesoderm development. The time-course comprised eight overlapping time-points spanning half of embryogenesis and captured a continuum of regulatory transitions as cells move from multipotency to different developmental lineages. I used this dataset to reconstruct developmental trajectories of all major cell-types, and uncover both the TFs and enhancers involved.

I then present two approaches that integrate the single-cell resolution of sci-ATAC-seq with perturbations of TFs and enhancers, as a means to dissect regulatory networks. First, I perturbed regulatory networks in *trans*, by mutating four key developmental TFs that drive *Drosophila* mesoderm development and used sci-ATAC-seq to jointly assess the regulatory outcome at both the cellular and molecular level. I demonstrate that this approach not only recovers previously described high-level phenotypes, but also more subtle alterations in cell fate, while simultaneously providing information on the TF's input at enhancers. Second, I perturbed regulatory networks in *cis* by exploiting natural sequence variation as a means for large-scale enhancer disruption. By profiling chromatin accessibility in hybrid embryos obtained from mating genetically diverse parent lines, I show how sci-ATAC-seq can be used to discover the cellular context affected by genetic variants and I discuss upcoming *in vivo* experiments to assess their impact on enhancer activity.

Zusammenfassung

Genregulatorische Netzwerke kontrollieren die genaue räumliche und zeitliche Ausführung von Entwicklungs-Genexpressionsprogrammen während der Embryogenese. Enhancer sind dabei Kernkomponenten dieser regulatorischen Netzwerke und dienen als Integrationsplattform für Entwicklungssignale, die von Transkriptionsfaktoren vermittelt werden. Das Entschlüsseln der Funktion von genregulatorischen Netzwerken und ihrer zugrundeliegenden Komponenten ist ein großes Ziel der modernen Biologie. In dieser Doktorarbeit nutze ich die Embryogenese von *Drosophila melanogaster* als gut untersuchtes Modelsystem und die jüngsten Fortschritte der Einzelzell-Genomik-Methoden, um die Auswirkungen von Störungen auf die entwicklungsregulatorischen Netzwerke zu untersuchen.

Ich begann mit der Optimierung eines Protokolls zur Profilierung der Zugänglichkeit des Chromatins in Einzelzellen durch sci-ATAC-seq in *Drosophila* Embryonen und wendete dieses an, um eine engmaschige Zeitreihe von über 20 000 Zellen während der Mesodermentwicklung herauszuarbeiten. Die Zeitreihe umfasste 8 überlappende Zeitpunkte, welche die Hälfte der Embryogenese abdeckten und hält regulatorische Übergänge fest, durch welche sich multipotente Zellen zu verschiedenen differenzierten Zelltypen entwickelten. Ich nutzte dieses Datenset, um die Entwicklungsverläufe der wichtigsten Zelltypen nachzubilden und die beteiligten Transkriptionsfaktoren und Enhancer zu identifizieren.

Ich präsentiere zwei Ansätze, welche die Einzelzell-Auflösung der sci-ATAC-seq Daten nutzen und zusammen mit Störungen von Transkriptionsfaktoren und Enhancern helfen regulatorische Netzwerke zu analysieren. Zunächst habe ich durch die Mutation von vier Schlüssel-Transkriptionsfaktoren, welche die Mesodermentwicklung in Drosophila antreiben, regulatorische Netzwerke in *trans* gestört. Ich habe dann sci-ATAC-seq genutzt, um die regulatorischen Auswirkungen sowohl auf zellulärer als auch auf molekularer Ebene gesamtheitlich zu untersuchen. Ich zeige auf, dass dieser Ansatz nicht nur zuvor beschriebene bedeutende Phänotypen rekapituliert, sondern auch subtilere Veränderungen des Zellschicksals nebst der Gewinnung von Erkenntnissen über die Rolle von Transkriptionsfaktoren an Enhancern. Zusätzlich habe ich natürliche Sequenzvariation benutzt, um Enhancern

im großen Format zu stören und so regulatorische Netzwerke in *cis* manipuliert. Durch die Profilierung der Chromatin-Zugänglichkeit in hybriden Embryonen, entstanden durch die Kreuzung genetisch diverser Elternlinien, zeige ich, wie sci-ATAC-seq genutzt werden kann, um den zellulären Kontext zu entdecken, der durch genetische Varianten beeinflusst wird. Ebenso erörtere ich künftige *in vivo* Experimente, um ihren Einfluss auf die Enhanceraktivität zu bewerten.

Table of Contents

List of Figures

List of Abbreviations

ATAC-seq	Assay for Transposase-Accessible Chromatin using sequencing
ChIP	Chromatin ImmunoPrecipitation
CRM	*Cis*-Regulatory Module
DHS	DNase Hypersensitive Site
FACS	Fluorescence-Activated Cell Sorting
GRN	Gene Regulatory Network
PTM	Post-Translational Modification
TF	Transcription Factor

1 Introduction

1.1 Embryonic development: from a single cell to a complex organism

A striking feature of multicellular development is the capacity of a single embryonic cell to generate a whole organism that comprises millions of cells with highly diversified and specialized functions. How does this morphological and functional diversity arise from the genome of a single cell? One early hypothesis was that as cells progress through development and become more diverse, their DNA also progressively changes and only the information required for a specific cellular identity is maintained. This hypothesis was rejected by seminal experiments in animal cloning (Gurdon et al., 1975; Wilmut et al., 1997), which succeeded in regenerating a whole animal using the genome of an adult differentiated cell, thereby disproving that genetic material is lost or substantially changed during development. The advent of high-throughput whole-genome sequencing has further confirmed that, with very few exceptions, all cells in an organism have the same genome, although with somatic variations. Therefore, cellular identity must be acquired by selectively activating and repressing the usage of a common genome, which we now know occurs through the regulation of gene expression programs over space and time as cells progress through development.

1.2 Gene regulatory networks drive cell fate decisions in embryonic development

Cellular identity is acquired during development by progressively transitioning through states with spatially and temporally restricted patterns of gene expression. The establishment of cellular identities needs to be tightly regulated in order to achieve the complexity and high degree of structure of animal tissues. Additionally, there must be control systems that ensure identities are consistently maintained throughout the organism's lifetime. The actualization of a cell's correct identity is orchestrated by highly sophisticated and interconnected transcriptional programs acting at specific stages, termed gene regulatory networks (GRNs) (Levine and Davidson, 2005).

1.2.1 The core components of regulatory networks: transcription factors and *cis*-regulatory modules

The gene regulatory networks that control development are mainly enacted through the action of a key group of regulatory proteins, transcription factors (TFs), at specific times and locations in embryos. Transcription factors are DNA-binding proteins that can activate, modulate and repress the transcription of genes. Their sequence-specific DNA binding targets them to *cis*-regulatory modules within the regulatory landscape of specific genes, in particular effector proteins that contribute to cellular identities (for example contractile proteins in muscle cells) and components of cell signaling pathways that are responsible for relaying developmental signals. TFs thereby specify and maintain cellular identity, giving a cell or tissue much of its morphological and functional characteristics (Reiter et al., 2017; Spitz and Furlong, 2012). A TF's expression is often shaped by both intrinsic clues, for example the presence of maternally deposited factors or other upstream TFs, and extrinsic environmental signals, such as cell signaling cascades, which give positional cues within the embryo. Most TFs thereby act in combinations to regulate a specific expression pattern at a specific stage of embryogenesis, and can work together with a different combination of factors at other stages or tissues. However, in some cases, expression of even a single transcription factor can be sufficient to drive cells to the acquisition of a distinct fate. For example in the fruit fly *Drosophila melanogaster*, the master regulator Twist, a basic helix-loop-helix (bHLH) transcription factor required for gastrulation and specification of the mesoderm, is sufficient to convert ectodermal cells to a mesodermal fate when ectopically expressed (Baylies and Bate, 1996). Similarly, its functional ortholog MyoD, is sufficient to convert fibroblast to a myogenic cell fate (Weintraub et al., 1989).

A second core component of gene regulatory networks are DNA sequences termed *cis*-regulatory modules (CRMs), through which transcription factors act (Wittkopp and Kalay, 2012). These elements are usually a few hundred base pair in size and have the capacity to regulate the transcriptional activity of genes. The *cis*- prefix is used to indicate that they regulate genes located on the same chromosome, in contrast to transcription factors which act in *trans* and can regulate genes farther away or on different chromosomes. CRMs make up a substantial portion of the genome as in both *Drosophila* and humans they are estimated to be in the range of hundreds of

thousands, massively outnumbering the ~ 18-25,000 protein-coding genes, a fact that reflects the necessity for complexity in gene regulation. CRMs are categorized into several classes depending on their sequence features and function, including gene promoters, however most of the tissue-specific regulation during developmental transitions takes place at a distinct class of regulatory elements, called enhancers (Farley et al., 2016; Shlyueva et al., 2014; Wittkopp and Kalay, 2012). DNA sequences are traditionally defined to be enhancers if they can enhance gene expression independently of distance and orientation (Banerji et al., 1981). Unlike promoter elements, enhancers can act over large distances, up to kilobases and even megabases from the gene they regulate (Furlong and Levine, 2018). The functional role of enhancers is to dictate the spatiotemporal expression of genes based on the activity imposed by TFs, which bind to DNA recognition motifs within enhancers. From a network perspective, enhancers can be considered as regulatory hubs that perform integrated signal processing of TF inputs. A single enhancer can display multiple binding sites for different TFs, and the recruited TFs can work cooperatively to modulate gene expression, which allows the delivery of more complex inputs. A gene's expression is usually regulated through the action of multiple enhancers; therefore, the expression pattern is a composite of activity directed by individual enhancers. The intricate combinatorial patterns of TF binding and integration between regulatory elements give rise to the extreme complexity of outcomes observed in the development of multicellular organisms.

1.2.2 The *Drosophila* mesoderm: a great model to study developmental regulatory networks

For many decades the fruit fly *Drosophila melanogaster* has been, and remains today, one of the most studied model organisms for both developmental biology and transcription/chromatin biology. This model offers several case studies for the principles of regulatory networks described above. *Drosophila* development begins with a fertilized embryo, which, like most animals, undergoes gastrulation (approx. 2-4 hours post fertilization) to form three germ layers: the mesoderm, endoderm and ectoderm. Within each germ layer, cell fate specification and terminal differentiation lead to a large variety of highly specialized cell types (Alberts et al., 2002). The mesoderm is the germ layer that gives rise to the muscle system from flies to humans,

and the key TFs regulating the subdivision into different muscle lineages are known and highly conserved among species. In *Drosophila melanogaster*, the mesoderm starts off as a relatively uniform monolayer of cells on the ventral side of the embryo. After gastrulation, the mesoderm divides and migrates dorsally, while at the same time becoming specified into diverse myogenic as well as non-myogenic lineages, such as the fat body. The muscle lineages include the somatic (striated) muscle, which is necessary for locomotion, the visceral (smooth) muscle, which envelopes the gut, and the cardiomyocytes, which later form the heart of the embryo. The gene regulatory network governing this process has been extensively characterized and modelled (Azpiazu and Frasch, 1993; Azpiazu et al., 1996; Bonn and Furlong, 2008; Mbodj et al., 2016; Wilczyski and Furlong, 2010; Yin and Frasch, 1998; Zaffran et al., 2001).

This process is driven by a hierarchical regulatory network of transcription factors (Figure 1a). Seminal genetic studies in *Drosophila* discovered that all of these factors are functionally required for the development of different muscle lineages, and their mutation leads to very drastic embryonic phenotypes such as missing or abnormal muscles. The master regulator *twist* sits at the top of the network and kickstarts the whole process by inducing the expression of *Mef2* (*Myocyte Enhancer Factor 2*) and *tinman* (*tin*; *Nkx2-5* homolog). *Mef2* is a pan-muscle regulator, expressed continuously from progenitors to differentiated muscle cells and is essential for myoblast fusion and terminal differentiation of all muscle types. Concordantly, *Mef2* mutants fail to form fully differentiated and functional muscle fibers (Bour et al., 1995). *Tinman* is essential for the subdivision of the dorsal mesoderm into the dorsal somatic muscle, the cardiogenic and visceral mesoderm; *tinman* mutants display a striking loss of all these dorsal mesoderm derivatives, in particular they have no heart or gut muscle (Azpiazu and Frasch, 1993).

Consistent with the role of *tinman* in the development of several muscle lineages, its spatiotemporal expression is very dynamic and under the strict control of four distinct enhancer elements (Figure 1b). Following gastrulation, Twist binds to recognition sequences within an enhancer in the gene body of *tinman*, termed enhancer B, which activates *tinman* and directs its broad expression throughout the trunk mesoderm. At this stage, *tinman* is also expressed in the embryonic head mesoderm, a pattern that is dictated by a second enhancer element, termed enhancer A. In the trunk mesoderm,

the regulation of *tinman* expression switches to being directed by the downstream enhancer D to specify the dorsal mesoderm. As the dorsal mesoderm diverges into visceral and cardiac mesoderm, *tinman* expression becomes solely restricted to the cardioblasts that will form the embryonic heart, under the control of a second downstream enhancer element, termed enhancer C (Yin et al., 1997).

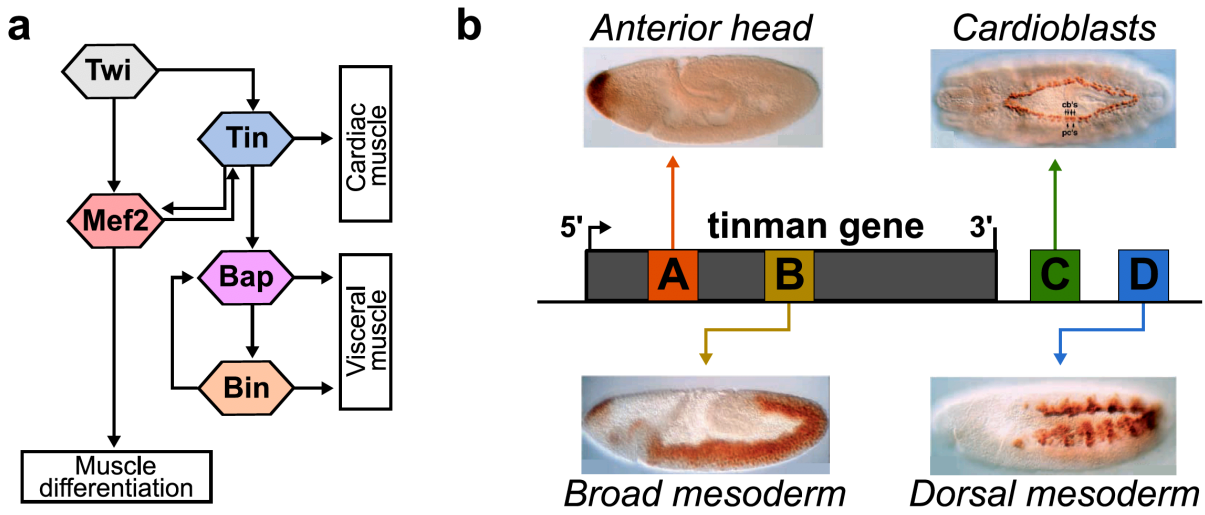

Figure 1. Mesoderm development is tightly regulated by a gene regulatory network of enhancers and transcription factors.

(a) Simplified schematic of the mesoderm gene regulatory network. Major lineage determining transcription factors identified by previous genetic studies are included. Adapted from (Zinzen et al., 2009) with permission License Number: 5118140343137.

(b) Schematic of the *tinman* gene locus and its enhancer elements (A-D). The embryo stainings depict the expression pattern driven by each enhancer element. Adapted from (Yin et al., 1997) under Attribution 4.0 International (CC BY 4.0).

More downstream in the mesoderm GRN network, Tinman regulates visceral mesoderm development by the activation of *bagpipe* (*bap;* NKx3-2 homolog) in progenitor cells, which in turn quickly initiates *biniou* (*bin*; FoxF2 homolog) expression and the onset of the visceral muscle lineage. Loss of either *bagpipe* or *biniou* leads to the loss of the circular visceral musculature that surrounds the gut (Azpiazu and Frasch, 1993; Zaffran et al., 2001). Each of these key TFs in turn regulates hundreds of downstream effector genes; for example, *Mef2* binds to roughly 1,000 genomic

regions and regulates the expression of ~200 direct target genes (Sandmann et al., 2006), such as contractile muscle proteins, granting each muscle sub-type their specific properties.

1.3 Genome-wide approaches for the dissection of regulatory networks

Dissecting gene regulatory networks and understanding how they drive the developmental progression of different cell lineages and tissues' development is a major goal of modern biology. A systems-level understanding of gene regulatory networks requires two key types of information. The first is the precise identification of all the regulatory elements (enhancers) within the network and the characterization of their activity states through space and time. The second is a detailed comprehension of how regulatory changes are induced – i.e. what is the functional input of these enhancers for their target gene's expression, and how is the input of TFs acting through these enhancers translated to regulatory output. Traditional genetic studies have dissected the main transcription factors and their target regulatory regions for a handful of regulatory networks, such as the *Drosophila melanogaster* mesoderm described above. However, the low throughput and time cost of these approaches has limited the extent to which regulatory networks can be deciphered, even for heavily studied model systems like *Drosophila*. The advent of methods that exploit the high throughput capacity of DNA sequencing has led to the ability to massively identify new regulatory elements and simultaneously track their activity genome-wide.

1.3.1 Monitoring transcription factor binding by ChIP-seq

As described above, defining the regulatory regions that are subjected to the action of specific transcription factors is crucial for deciphering regulatory networks. Chromatin immunoprecipitation (ChIP) enables the detection of protein-DNA binding *in vitro* and *in vivo,* by using an antibody that specifically recognizes the factor of interest as a means to recover the bound DNA fragments. Immunoprecipitating protein:DNA complexes and then sequencing the isolated DNA, an approach termed ChIP-seq, reveals the genome-wide occupancy of the target binding protein (Park, 2009). This method can therefore detect the repertoire of *cis*-regulatory elements (enhancers)

acted upon by TFs at a given time and it has been particularly powerful in revealing the temporal dynamic patterns of TF activity during embryonic development (Junion et al., 2012; Wilczyski and Furlong, 2010; Zinzen et al., 2009). ChIP-seq has been used extensively to reconstruct the binding maps of hundreds of TFs in a large collection of species, including bacteria (Galagan et al., 2013), flies (Kudron et al., 2018) and humans (Dunham et al., 2012).

The major limitation of ChIP is that it requires the generation of highly specific antibodies that recognize the protein of interest. While major efforts have been undertaken to improve antibody production and to assess their quality (Egelhofer et al., 2011), issues with reproducibility between antibodies and unavailability of suitable ones for certain targets are still common. Another important aspect to consider is that occupancy does not necessarily imply functionality; therefore, the role of a bound factor at a given region and the effect on the expression of predicted target genes still needs to be validated in downstream activity assays.

1.3.2 Discovery and annotation of regulatory regions by histone PTMs profiling

In order to fit inside the tiny nuclear space, eukaryotic genomes are packed into chromatin, a highly structured assembly of DNA and proteins. The fundamental packaging unit of chromatin is the nucleosome, a structure consisting of ~147 base pairs of DNA wrapped around an octamer of proteins called histones. Residues on the terminal tails of histone proteins harbor a plethora of distinct post-translational chemical modifications (Kouzarides, 2007), including acetylation, methylation, phosphorylation and ubiquitination, to name just a few. Histone post-translational modifications (PTMs) can influence chromatin state and gene expression in two ways: (I) modifications that are charged can alter the overall strength of the DNA-histone interaction and thus promote relaxation or condensation of the chromatin (Dhall et al., 2014; Fenley et al., 2018; Otterstrom et al., 2019), (II) modifications can provide docking sites that are specifically recognized by protein complexes, including chromatin remodelers, which act to modify the chromatin structure directly or further recruit other effector proteins to regulatory regions (Margueron et al., 2005; Nightingale et al., 2006).

Genome-wide profiling of histone PTMs by ChIP-seq revealed that distinct modifications are associated with distinct genomic regions and correlate with their activity state. For example, acetylation of lysine 27 on histone 3 (H3K27ac), is enriched at promoters of active genes and is a mark of active enhancers (Bonn et al., 2012a; Creyghton et al., 2010). Mono-methylation of lysine 4 on histone 3 (H3K4me1) is mostly enriched downstream of the transcription start sites of active genes and it marks both weak and strong enhancers (Barski et al., 2007; Heintzman et al., 2007) as well as enhancers in a repressed or inactive state (Bonn et al., 2012a). Therefore, although H3K4me1 is a good indicator of enhancers, it is not specific to a specific activity state (Bonn et al., 2012a). Unlike mono-methylation, tri-methylation of K4 (H3K4me3) is not present at enhancers, but resides at the transcription start site (TSS) of active genes (Guenther et al., 2007; Heintzman et al., 2007). Tri-methylation of lysine 27 on histone 3 (H3K27me3) is instead associated with heterochromatin and polycomb silenced regions (Barski et al., 2007). Because of these associations between histone PTMs and genomic activity, ChIP-seq profiling of histone marks has become a mainstream approach to annotate non-coding regions, in particular to identify novel regulatory elements, and to predict the biological activity state of chromatin regions (Ernst et al., 2011).

1.3.3 Chromatin accessibility profiling as a tool to track regulatory events

As described in the previous section, chromatin is a dynamic structure and the degree of tightness and relaxation in its compaction is crucial for cellular processes (Kornberg and Lorch, 1992; Li and Reinberg, 2011). In particular, the physical accessibility of genomic regions is a key feature in their regulation; active regulatory elements are generally depleted of histones and in an open chromatin configuration that is permissive to the binding of transcription factors (Lee et al., 2004). This simple fact implies that profiling the degree of chromatin accessibility of genomic regions can be exploited to discover new potential regulatory elements and track their occupancy status *in vivo*. Compared to ChIP, this approach does not require knowledge of the target beforehand and avoids potential biases introduced by the use of antibodies. Moreover, unlike histone PTMs, chromatin accessibility is a stereotypical feature of all active CRMs, and thus it should enable the identification of the entire repertoire of regulatory elements in use at a specific time in a given cell-type or cell state.

Chromatin accessibility can be probed genome-wide by digestion with DNaseI nuclease (deoxyribonuclease I hypersensitive sites sequencing (DNase-seq)) (Boyle et al., 2008), physical isolation of protein-bound and protein-free DNA fragments (formaldehyde-assisted isolation of regulatory elements sequencing (FAIRE-seq)) (Giresi et al., 2007) or by Tn5 transposition (assay for transposase-accessible chromatin using sequencing (ATAC-seq)) (Buenrostro et al., 2013). Among these methods, ATAC-seq has emerged as the most widely used, due to its simple and robust protocol and high sensitivity, allowing it to be applied to low cell numbers (Minnoye et al., 2021). ATAC-seq relies on the use of a hyperactive prokaryotic Tn5 transposase, which inserts itself into the accessible chromatin regions of the genome and tags these sites with sequencing compatible adaptors, allowing to recover genome-wide chromatin accessibility information (Buenrostro et al., 2013) (Figure 2). Over the past decade the use of ATAC-seq and other chromatin accessibility profiling methods has produced atlases of regulatory elements and their activity state in a diverse range of primary cells, organs and organisms (Dunham et al., 2012; modENCODE Consortium et al., 2011; Reddington et al., 2020; Roadmap Epigenomics Consortium et al., 2015; Thurman et al., 2012).

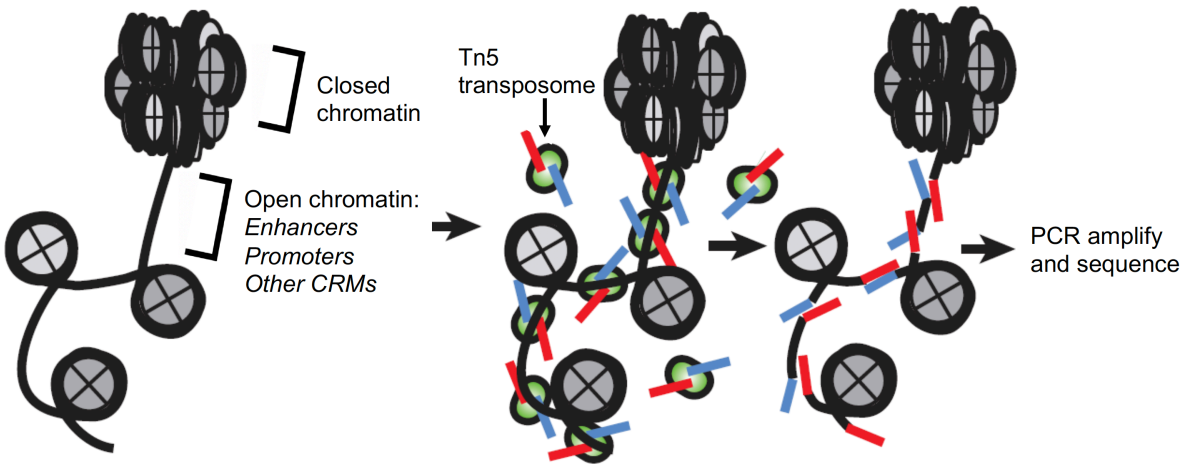

Figure 2. Chromatin accessibility profiling by ATAC-seq.

Regulatory elements such as enhancers and promoters are in an open chromatin configuration when active. Tn5 transposase (green) inserts itself into accessible chromatin regions of the genome and tags them with sequencing-ready adapters (blue and red). These regions can be

directly PCR amplified and sequenced to reveal their genomic position. Chromatin regions in a closed conformation are not accessible to Tn5 and therefore they are not targeted. Adapted from (Buenrostro et al., 2013) with permission License Number: 5118121364313.

1.3.4 Chromatin accessibility as a predictor of enhancer spatiotemporal activity

As chromatin accessibility is reflective of regulatory elements usage, it is, in many instances, cell-type and cell-state dependent. This is particularly true during development, when the accessibility landscape is highly diverse and becomes progressively more different as lineages diverge (Bozek et al., 2019; Daugherty et al., 2017; McKay and Lieb, 2013; Reddington et al., 2020; Thomas et al., 2011; Yan et al., 2019). In accordance with enhancers being the primary drivers of tissue-specific gene expression, accessibility is very dynamic at distal regulatory elements across cell-types, while accessibility at promoters is more stable (Heintzman et al., 2009; Reddington et al., 2020; Thurman et al., 2012).

For this reason, chromatin accessibility has been used to identity novel enhancers and to predict their spatiotemporal activity (Kwasnieski et al., 2014; Shlyueva et al., 2014). While enhancer accessibility generally correlates well with activity, this association is not perfect. There are in fact many examples of enhancers that display cell-type restricted activity while being simultaneously accessible in two or more lineages (Bozek et al., 2019; Reddington et al., 2020; Shashikant et al., 2018). In these instances, chromatin accessibility at the regulatory element is clearly not sufficient to drive enhancer activation. Accessibility at these regions might be due to the binding of ubiquitous factors or to the binding of a repressor that is blocking activation (Uyehara et al., 2017). Alternatively, accessibility might reflect enhancers that are primed for future activation (Liber et al., 2010; Mercer et al., 2011).

This highlights the importance of integrating additional activity signatures in enhancer spatiotemporal predictions, including histone PTMs (H3K27ac), TF and polymerase occupancy and enhancer transcription (Bonn et al., 2012a; Mikhaylichenko et al., 2018; Zinzen et al., 2009). Nevertheless, chromatin accessibility still remains a powerful predictor of enhancer spatiotemporal activity, as testified by the fact that the correct activity pattern of enhancers accessible in more than one cell-type can often be

identified based on the quantitative intensity of the accessibility signal (Bozek et al., 2019; Reddington et al., 2020; Shashikant et al., 2018). Moreover, recent single-cell studies in the *Drosophila* embryo (Cusanovich et al., 2018a) and eye imaginal disc (Bravo González-Blas et al., 2020), remarkably showed that cell-type specific chromatin accessibility alone is sufficient to predict *in vivo* enhancer activity in the majority of cases tested (>80%) and to identify cell-types.

1.3.5 Single-cell methods allow to resolve cell-type specific regulatory events

The efforts described above revealed that chromatin accessibility in distal regulatory sites (enhancers) is tightly linked to the occupancy of regulatory elements and is therefore highly cell-type specific. Crucially, the heterogeneity of cells present in developing embryos poses a major obstacle to the application of genomic approaches such as ATAC-seq. Typically, tens of thousands of cells are used as input for the assay. While this is much better than the millions of cells needed for DNase-seq, the resulting signal from bulk ATAC-seq still represents an ensemble average of the cell types and states profiled. This complication acts to limit the detection of the dynamics of cell-type specific regulatory networks and in turn, our capacity to decode the relationship between activity of individual elements and their sequence features and other properties during embryogenesis.

An approach to get around this problem consists in physically isolating the tissues/cell-types of interest, either by tissue dissection or Fluorescence-Activated Cell Sorting (FACS). While tissue dissection has for example been applied to study fly (Haines and Eisen, 2018; McKay and Lieb, 2013) and mouse (Blow et al., 2010; Soshnikova and Duboule, 2009) development, it remains challenging to perform, especially at earlier time points when tissues have not yet formed, and its resolution is coarse-grained, as it is often not possible to isolate distinct cell-types within a tissue. On the other hand, FACS can isolate cell-types from complex heterogeneous mixtures based on light-scattering properties and fluorescence emission (Adan et al., 2017). This methodology has been particularly powerful when applied to immunobiology, as samples are easy to obtain and cell-types are readily identifiable with a wide panel of specific surface markers. Protocols for the efficient cell labelling and sorting of cells in embryonic

samples have been implemented, thus enabling genomic profiling of tissues and cell-types during embryogenesis (Bonn et al., 2012b; Reddington et al., 2020). Nevertheless, FACS optimization for embryo samples remains laborious, and often it is not possible to find specific markers and antibodies that can discriminate cellular sub-populations or rapid developmental transitions.

The ideal approach would consist in measuring each single cell in a whole sample so that the full spectrum of cellular diversity and molecular variation can be captured. Following this ideal, a lot of effort has been dedicated in recent years to the development of methods to probe the transcriptome and chromatin state in single cells (Kelsey and Stegle, 2017). Recently developed protocols for single-cell ATAC-sequencing (scATAC-seq) allow the unbiased identification of cell-type specific regulatory elements from mixtures of heterogeneous populations with a throughput of tens of thousands of cells per assay (Cusanovich et al., 2015; Minnoye et al., 2021) (Figure 3). scATAC-seq has already been successfully applied to resolve the regulatory landscape of adult mouse tissues (Cusanovich et al., 2018b), human and mouse brain (Lake et al., 2018; Preissl et al., 2018; Sinnamon et al., 2019), tumor biopsies (Satpathy et al., 2019), maize (Marand et al., 2021) and *Drosophila* embryos (Cusanovich et al., 2018a), plus many others.

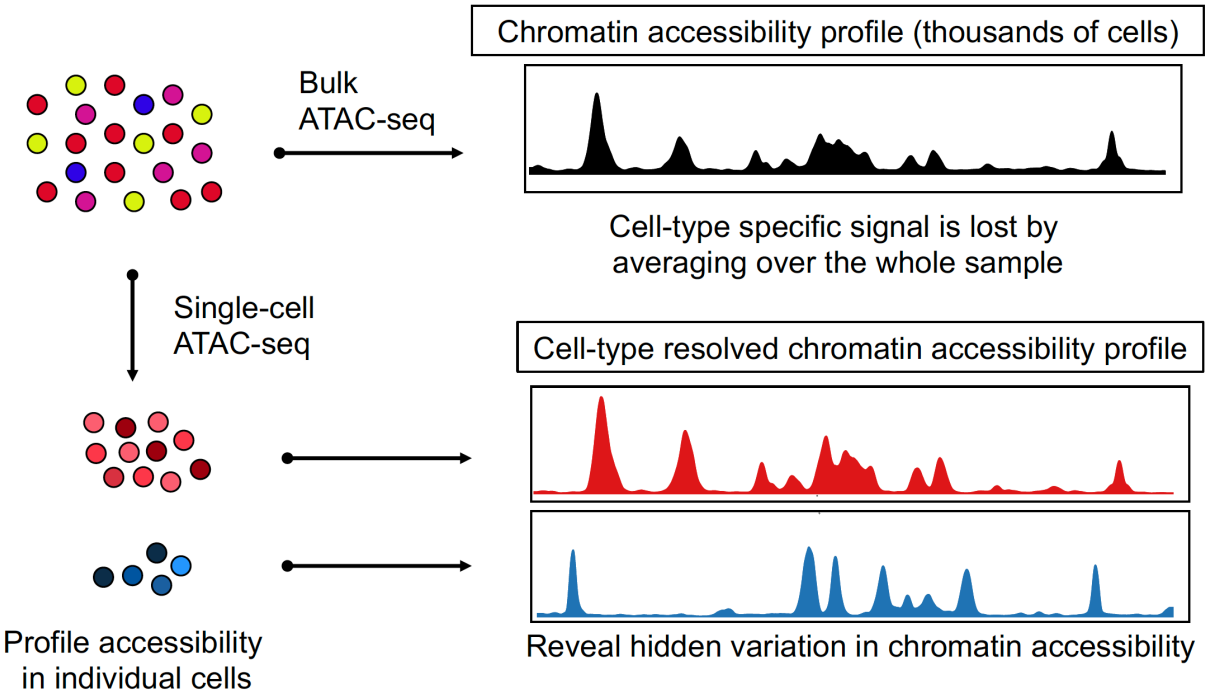

Figure 3. Single-cell ATAC-seq resolves chromatin accessibility heterogenicity in complex samples.

When performing ATAC-seq profiling of complex samples, such as developing embryos, cell-type specific signals are masked by signal averaging over the whole population heterogenicity. This limitation can be overcome by profiling chromatin accessibility in single cells, which allows to recover cell-type specific information.

Among emerging single-cell technologies, transcriptome profiling by single-cell RNA sequencing (scRNA-seq) has been widely employed to disentangle expression heterogenicity in complex samples and has also been applied to resolve cellular diversity and progression during development, most notably in early *Drosophila* embryos (Karaiskos et al., 2017), *C. elegans* (Tintori et al., 2016), zebrafish (Farrell et al., 2018) and mouse organogenesis (Cao et al., 2019). While the output generated by scRNA-seq and scATAC-seq might seem similar at first glance, they are really not equivalent and thoughtful consideration needs to be given to decide the most suitable method for a given study. Single-cell RNA-seq is very powerful in resolving the complexity of cell-types present in heterogenous samples and to characterize their gene expression programs, however it fails to reveal how the observed expression changes are enacted by the regulatory genome. On the other hand, scATAC-seq is

much more suited to connect the regulatory genome with its functional output. Besides enabling the capture of the regulatory landscape, scATAC-seq allows to identify novel tissue-specific enhancers (Bravo González-Blas et al., 2020; Cusanovich et al., 2018a), identify the motif grammar and the TFs acting on such enhancers (Cusanovich et al., 2018b; Preissl et al., 2018) and link regulatory elements to their target genes (Pliner et al., 2018). It can also give a rough proxy for gene expression, as I show in my data below. Therefore scATAC-seq should be the method of choice when the scope of the study is to probe the regulatory genome and dissect the underlying regulatory networks.

1.3.6 Single-cell chromatin accessibility profiling by scATAC-seq

Two methods were developed independently and simultaneously for single-cell ATAC-seq, one based on droplet microfluidics (Buenrostro et al., 2015) and one based on combinatorial indexing (Cusanovich et al., 2015) (both reviewed in (Pott and Lieb, 2015) and (Minnoye et al., 2021)). The scATAC-seq protocol from (Buenrostro et al., 2015) relies on physical isolation of single cells. A programmable microfluidics device is used to compartmentalize single cells into nanoscale reaction chambers and, after cell viability is confirmed with the use of a microscope, ATAC-seq is performed on each captured cell individually. Given the highly controlled reaction environment, the tagmentation reaction tends to be very efficient with this protocol, while the major downside is scalability, as only a few hundred cells can be processed at a time. Instead, the scATAC-seq protocol from (Cusanovich et al., 2015) gets around the challenge of physically isolating single cells by performing subsequent rounds of combinatorial indexing to uniquely barcode individual cells, an approach that has been named single-cell ATAC-seq by combinatorial indexing or sci-ATAC-seq. Initially, a cell sorter is used to dispense a defined number of nuclei (typically 2500) into the wells of a 96-well plate, each containing a Tn5 transposase loaded with a unique combination of barcodes (Figure 4a). The tagmentation reaction thus introduces the first round of indexes. The nuclei are then pooled, mixed and dispensed in each well of 96-well plates in the defined number of 25 nuclei. Each well contains a unique combination of barcoded adaptors, which are integrated during the PCR amplification thus providing the second round of indexing. After sequencing, it is then possible to identify each cell

based on a unique set of barcodes. Because the number of nuclei aliquoted in each well for the second indexing round is at least 100 times lower than the first, the number of collisions, i.e. cells having identical barcodes, is statistically ensured to be low (estimated and measured at ~ 11%) (Cusanovich et al., 2015). While Tn5 tagmentation seems to be less efficient for sci-ATAC-seq compared to microfluidics, this protocol has drastically higher throughput, in the order of tens of thousands of cells in a single run, and routinely obtains a median of 20,000 reads per cell with my optimizations (discussed in Chapter 3). A recently developed protocol for 3-level sci-ATAC-seq (sci-ATAC-seq3) enables to further scale up the assay throughput, up to a million cells, by increasing the number of combinatorial indexing rounds (Domcke et al., 2020).

Despite the differences in approaches, the output is the same: for both assays the sites of Tn5 integration, representing accessible chromatin sites of the genome, are recovered for each single cell. This is usually converted into a matrix that lists for each genomic site the count of insertions per cell (Figure 4b). Because diploid genomes have a maximum of two homologous chromosomes that regulatory elements can be measured from, the count matrix is almost binary and very sparse. A first analysis step usually consists in clustering the count matrix to reveal the dynamics in the regulatory landscape and the cell-type diversity present in the dataset (Figure 4b). Clustering and downstream analysis are performed with computational methods that have been developed ad-hoc to deal with the challenging nature of scATAC-seq data (Bravo González-Blas et al., 2019; Granja et al., 2021; Minnoye et al., 2021).

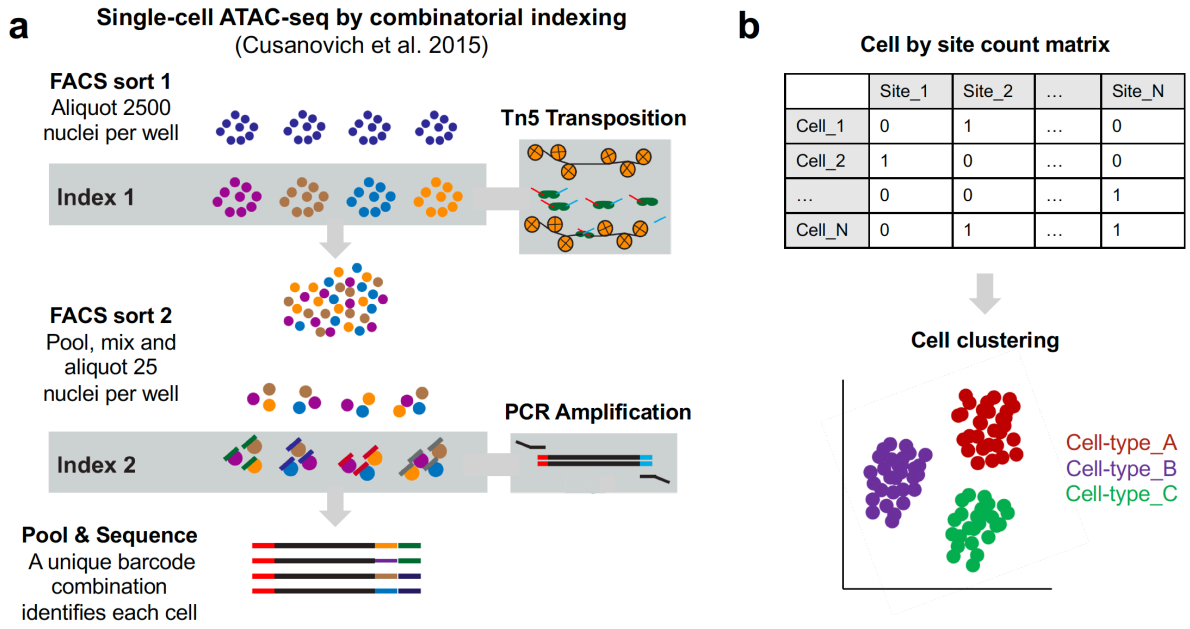

Figure 4. Illustration of single-cell ATAC-seq by combinatorial indexing.

(a) Simplified schematic of the strategy for profiling single-cell chromatin accessibility by combinatorial indexing. Adapted from (Pott and Lieb, 2015) with permission granted under Creative Commons CC BY license.

(b) The output of scATAC-seq methods is essentially a count matrix of Tn5 insertions per cell. A typical analysis procedure consists in cluster the count matrix to reveal the dynamics in the regulatory landscape and the cell-type diversity present in the dataset.

At the beginning of my PhD, sci-ATAC-seq was the only method available having the necessary throughput to profile a full developmental time course. Recent advances have increased both the throughput and the availability of droplet microfluidics. For example, commercial kits from 10x Genomics (Chromium Next Gem Single Cell ATAC-seq Library Kit) (Satpathy et al., 2019) enable scATAC-seq profiling at a similar throughout as sci-ATAC-seq, although at a significant cost. The line separating the two methods has also been blurred, with the implementation of droplet-based combinatorial indexing approaches (Lareau et al., 2019), which incorporate the advantages of both. A particularly rapid area of new development has been the integration of scATAC-seq with other single cell measurements, enabling the multi-omic profiling of several molecular layers from the same cell, including the simultaneous capture of chromatin accessibility and either the transcriptome (Cao et

al., 2018; Ma et al., 2020), protein-abundance (Chen et al., 2018) or DNA methylation (Argelaguet et al., 2019; Pott, 2016). This trend is set to continue to expand scATAC-seq capability and to integrate even more single-cell omics assays in the coming years, leading to new exciting avenues of future research.

1.3.7 Application of scATAC-seq to study embryonic development

Single-cell ATAC-seq can be extremely powerful when applied to embryonic development, as the samples are naturally very heterogenous cell mixtures undergoing rapid and transient changes and thus are challenging to tackle by conventional isolation-based methods, such as dissection or FACS. Our lab was the first to profile chromatin accessibility during embryonic development at single-cell resolution, by applying sci-ATAC-seq to over 20,000 cells spanning three major stages of *Drosophila melanogaster* embryogenesis (Cusanovich et al., 2018a). This seminal study successfully demonstrated that: a) the dynamics of regulatory element usage in single cells can be profiled through developmental time, b) cell identities can be identified based only on their differential usage of regulatory elements and c) single cell profiles of chromatin accessibility can identify and predict tissue-specific enhancer activity. Taken together, this suggests that sci-ATAC-seq can be used to reconstruct the progression of regulatory trajectories of single cells through time. More recent studies in mouse (Chung et al., 2019; Pijuan-Sala et al., 2019; Preissl et al., 2018) and human (Domcke et al., 2020; Ranzoni et al., 2021) have also demonstrated the use of scATAC-seq to resolve the chromatin landscape of cell-types during development, to find new regulatory elements active in each cell-type and to identify TFs acting on these elements. Therefore, scATAC-seq provides a direct approach to uncover enhancer usage in different tissues and stages of development and paves the way to the dissection of cell-type specific regulatory networks.

A limitation of these studies is that they are quite restricted in the temporal window that was assayed. For example, mice embryos were only assayed once (Pijuan-Sala et al., 2020) or with gaps of days in between time points (Preissl et al., 2018). Even for *Drosophila*, where development is very rapid, the three time points assayed in (Cusanovich et al., 2018a) are not continuous but separated by two-hour gaps, and in total they cover only 25% of embryogenesis (6/24 hours). In order to capture rapid

developmental transitions and reconstruct continuous regulatory trajectories, there is the need to perform more dense and comprehensive time courses.

Moreover, the new resolution provided by single-cell technologies should enable moving beyond descriptive cataloguing of cell-types and regulatory elements. In order to elucidate novel principles in the regulatory dynamics of embryonic lineage trajectories, mutations of regulatory elements and TFs should be introduced in the study system. A step in this direction has been taken with Perturb-ATAC (Rubin et al., 2019) and Spear-ATAC (Pierce et al., 2021), two methods that combine scATAC-seq with CRISPR screening to assess the impact of mutations in cell culture models. In model organisms, where large-scale CRISPR mutagenesis is not possible, mutations of key developmental TFs in *trans* and natural sequence variation to perturb regulatory elements in *cis* can be harnessed to introduce perturbations in regulatory networks and probe the functionality of its components for embryonic development. In my thesis I apply both of these approaches during embryonic development.

1.4 Thesis overview and aims

Over the past decade, substantial efforts have been dedicated to the production of atlases of regulatory elements and their activity state in a diverse range of primary cells, organs and organisms, as described above. The precise level of characterization afforded by single-cell technologies, and the scale at which they can be applied, offers new opportunities to decipher the regulatory networks that drive embryogenesis at unprecedented resolution, as demonstrated in several model organisms, from *Drosophila* to human (Argelaguet et al., 2019; Cusanovich et al., 2018a; Domcke et al., 2020; Pijuan-Sala et al., 2020).

However, none have so far combined this data with mutants for multiple transcription factors within a regulatory network to dissect the functional input of different regulators to lineage trajectories and chromatin dynamics. Understanding how regulatory networks are initiated and coordinated to achieve a specific output necessarily requires the integration of single-cell characterization with the capacity to selectively perturb components of the network. *Drosophila melanogaster* is a very powerful model in this

regard as it offers the possibility to perform precisely controlled genetic manipulations in the context of a well-defined developmental system that has been studied extensively, and can thereby serve as a model system to move beyond descriptive atlases.

During the course of my Ph.D. I developed three projects with the aim to characterize and functionally dissect regulatory networks during *Drosophila melanogaster* embryogenesis. To this end, I first worked to implement a protocol for single-cell chromatin accessibility profiling that is optimized for *Drosophila* embryos. I then applied this method in the context of transcription factor mutations and natural sequence variation, as means to perturb regulatory networks. Each project is presented in this thesis and described in a separate results chapter:

- Chapter 3: “Optimization of sci-ATAC-seq to profile single-cell chromatin accessibility in *Drosophila* embryos”. In this chapter, I describe the application of single-cell chromatin accessibility profiling by sci-ATAC-seq, using ‘home-made’ Tn5, and its optimization for *Drosophila* embryo samples. Major achievements include a considerable increase in the assay coverage per cell and a 10-fold reduction in reagents costs.

- Chapter 4: “Using sci-ATAC-seq to phenotype mutants at both a cellular and molecular level”. In this chapter, I first generated a high-resolution regulatory trajectory of mesoderm development using sci-ATAC-seq, from pluripotency, through specification and differentiation in wild-type embryos. I then combined this with sci-ATAC-seq after genetic perturbation of lineage specific transcription factors within the mesoderm regulatory network. Major achievements include the first regulatory atlas of *Drosophila* mesoderm development, which I generated at very high temporal resolution, the identification of thousands of new regulatory elements and many new transcription factors active in each lineage, and the characterization of the cellular and molecular phenotypes of four transcription factor mutants.

- Chapter 5: “Using sci-ATAC-seq to uncover cell-type specific genetic perturbations”. In this chapter, I demonstrate how single-cell chromatin

accessibility profiling can be used as a tool to discover the cellular context in which natural sequence variation impacts regulatory elements. Major achievements include the profiling of genetically diverse *Drosophila melanogaster* F1 embryos at multiple developmental time points and the establishment of a framework to map chromatin changes due to genetic imbalances to individual cell types.

2 Materials and Methods

2.1 Generation of transcription factor mutants by CRISPR

2.1.1 Rationale

Balancer chromosomes are highly scrambled chromosomes that prevent the recovery of recombinants by suppressing most of genetic recombination between homologous chromosomes during meiosis. They are typically homozygous lethal, and maintained in trans to a non-balancer homologous chromosome. When placed in trans to a recessive lethal mutation, the only embryos that can survive to adulthood are the trans-heterozygous mutation/balancer, with the homozygous mutation/mutation and balancer/balancer offspring being lethal. Recessive lethal mutations can thereby be maintained in trans to a balancer for decades (indefinitely). However, as any additional spurious mutations on the mutant of interest's chromosome also cannot recombine off the chromosome due to the presence of the balancer chromosome, old mutant stocks naturally accumulate other deleterious mutations. As the loss-of-function lines for all four mesodermal TFs assessed in Chapter 4 were generated twenty or more years ago, they will have accumulated many additional mutations, which are also maintained by the balancer chromosomes. In addition, as they were made by different labs at different stages, they also have different genetic backgrounds. I therefore decided to generate clean loss-of-function mutants for all four TFs in the same isogenic background.

2.1.2 CRISPR design

As the previous alleles were molecularly characterized and demonstrated to be loss-of-function, I used CRISPR induced template directed homology to regenerate the same loss-of-function alleles for each factor. Specifically, I regenerated the *Mef2* 22.21 allele (Flybase ID FBal0033789; (Bour et al., 1995)), *tinman* EC40 allele (Flybase ID FBal0032861; (Bodmer, 1993)) and *biniou* R22 allele (Flybase ID FBal0043738;

(Zaffran et al., 2001)). These are all single-nucleotide nonsense mutations that introduce a premature stop codon and are therefore protein nulls. As the gene *bagpipe* (Flybase ID FBgn0004862) does not have a characterized loss-of-function allele, the mutant phenotype was characterized using a deficiency, I applied the same CRISPR approach to introduce a nonsense mutation at 3R:G21389189T, which is located in the first exon and 180 bp before the TF's DNA binding domain. These mutations were introduced in a clean, isogenic and fully sequenced fly line that expresses Cas9 in the germline under the Vasa promoter w[1118]; PBac{y[+mDint2]=vas-Cas9}VK00027 (Bloomington stock 51324).

A single stranded oligonucleotide (ssODN) was designed for each locus (Table 1) to serve as a template for HDR following the Cas9 induced double strand break, based on the protocol available on the flyCRISPR website (https://flycrispr.org/protocols/ssodn/) (Gratz et al., 2015). The template ssODNs were designed to include additional features besides the intended single-nucleotide nonsense mutation: (1) a restriction site for SacI (GAGCTC; NEB) was introduced downstream the premature stop codon to be used for screening. (2) A thymine nucleotide was inserted immediately upstream the SacI restriction site, causing a frame-shift mutation that generates a second premature stop codon (TGA); this would serve to terminate translation in case of read-through after the first stop codon. (3) point mutations in the PAM or the gRNA seed to prevent re-cutting by Cas9. The ssODNs were synthesized by Integrated DNA Technologies (IDT). A single gRNA against each target locus (Table 2) was designed using flyCRISPR target finder (http://targetfinder.flycrispr.neuro.brown.edu).

Table 1. Single-stranded oligo donors sequence (5'-3').

Mef2	CAGCCGAGCCGAGCACAAGGACACCACCATCTCCGTTTCCATCGCAGGAG GATAGGAAATCTGTTGCCATGGGCCGCAAAAAAATTTAAATATCACGCATCA CCGATTGAGCTCAATCGGCAGGTGAGTGTCACAATGTGTTTGCCAATCCGT AGCTATAATAATACAAATTCGCTGAATATCTCTTGTTTGCTATCCTC
tinman	GGGGCTCGGACTTCAACTTCAGATGCTTGGCGATGCCCTCGCAGTCGATG TCGCCACGTTTCGATTTGTAGCGCCGATTCTGGAACCAGAGCTCACACTTA GGTGGCGGACAGGTTAAGCTTTTGCGCGATTATCTCGCGCTCCGCACCCG TCAGATACTTTTTGAGTCGAAAGCGACACTCCAGCTCCAGGACCTGTGC
biniou	CAACGGTTCCGAATCCTCGCCGCCCACTCAAAATCATCTGGAGAACAAGGT TTCGGGCTCAGCGGTGGTTGGTACAGGTGGTTCCAGTCAGTAGGATTGAG CTCAGCACTCCGGATACCACCAAGAAGTCTGGTACCCGGCGGCCAGAGAA ACCAGCACTCAGCTATATCAATATGATTGGACATGCGATCAAGGAATC
bagpipe	GTTGCTGGGTTGCCTGGCACTTGGTTGGATCTCCTTGGGTTGGGTCAGCTT GTAGAGGCCGAGATCTCGGCAGCAGAGTGGCGGCGACTTGGAGAGCTCAT CGCTATCGATCGGAGGAGGGTTTCAGCTTTTCCGGTTCCGGTTCGGAGTC CACGCTGGACATGCGACGTGTTTCCGGATTGCTGCGCGTTAGGATATCG

Table 2. Guide RNAs sequence (5'-3').

Mef2	Sense	CTTCGCATCACCGATGAACGCAAT
	Antisense	AAACATTGCGTTCATCGGTGATGC
tinman	Sense	CTTCGGCGCAAAAGCTTAACCTGT
	Antisense	AAACACAGGTTAAGCTTTTGCGCC
biniou	Sense	CTTCGTGGTGGTATCCGGAGTGCT
	Antisense	AAACAGCACTCCGGATACCACCAC
bagpipe	Sense	CTTCGGAGATCGATCGCTCGCGAT
	Antisense	AAACATCGCGAGCGATCGATCTCC

2.1.3 CRISPR lines generation

The gRNAs were cloned in the vector pBs-U6-gRNA-BbsI (Addgene #45946) and injected together with the ssODNs into embryos of the Vasa-Cas9 fly line described above. All embryo injections were performed by Alessandra Reversi of the EMBL *Drosophila* injection service. I crossed back emerging flies to the same isogenic Vasa-Cas9 line in trans to a sequenced balancer for chromosome 2 (If/Cyo, IsoVasCas9; *Mef2* allele) or chromosome 3 (IsoVasCas9, Sb/TM3 Ser; other alleles). Screening was performed by SacI restriction digestion of the PCR amplified locus. I confirmed that the correct mutant alleles were regenerated for each locus by three independent

methods; (1) the intended nonsense mutations were confirmed by Sanger sequencing, (2) by a genetic complementation test, which showed that the new alleles non-complement the lines carrying the “original” mutant alleles described above, as expected, and (3) immunostaining showed that the new alleles recapitulate the known mutant phenotypes.

2.2 *Drosophila* embryo collection

2.2.1 Embryo collection for the mesoderm sci-ATAC-seq datasets

Drosophila melanogaster embryos were collected and fixed as previously described (Bonn et al., 2012b; Sandmann et al., 2007). In summary, embryos were collected in staged two-hour windows following three one-hour pre-lays to clear the females and synchronize the collections, which were aged at 25°C to the corresponding time window (3-5 hr, 4-6 hr, 5-7 hr, 6-8 hr, 7-9 hr, 8-10 hr, 9-11 hr and 10-12 hr for wild-type collections, 10-12 hr for *Mef2*, 5-7 hr for *tinman* and 6-8 hr for *biniou* and *bagpipe* mutant embryo collections). Embryos were dechorionated in 50% bleach for 2 min and fixed with 1.8% formaldehyde for 15 min. For *Mef2* hand-sorted mutant embryo collections, homozygous mutant embryos (GFP negative) were hand-sorted from their GFP marker balancer siblings using the stock dMef222.21/CyO, twi-Gal4, UAS-GFP, under a dissection microscope prior to fixation. After 15 minutes of formaldehyde fixation, embryos were quenched with glycine, washed, dried, snap frozen in liquid nitrogen and stored at − 80 °C until use.

2.2.2 Embryo collection for the F1 sci-ATAC-seq datasets

Drosophila melanogaster F1 hybrid embryos were obtained by crossing females from a common maternal line with males from four genetically diverse inbred lines (DGRP-307, DGRP-639, DGRP-712, DGRP-852) from the *Drosophila melanogaster* genetic reference panel (DGRP) (Floc’hlay et al., 2021; MacKay et al., 2012). Embryos were collected in staged two-hour windows following three one-hour pre-lays to clear the females and synchronize the collections, which were aged at 25°C to the

corresponding time window (2-4 hr, 6-8 hr and 10-12 hr). Fixation and storage of the embryos was performed as described above for the mesoderm wild-type dataset. All fly crosses and F1 hybrid embryo collections described here were performed by Dr. Bingqing Zhao.

2.3 Nuclear isolation and preparation for sci-ATAC-seq

2.3.1 Nuclear isolation and quantification

Embryo dissociation and nuclear isolation were performed using a dounce homogenizer and a 20/22G needle as previously described (Bonn et al., 2012b). Briefly, frozen embryos were dounced with a glass homogenizer placed on ice, 20 times using a loose pestle and 10 times using a tight pestle. The homogenate was filtered through two layers of Miracloth (Millipore #475855) into a 15 mL tube and centrifuged at 3500 rpm for 3 min. The nuclei pellet was resuspended in 3 mL PBT (PBS + 0.1% TritonX-100) and the nuclei were dissociated using a 5 mL syringe by passing them ten times through a 20G needle and ten more times through a 22G needle. The homogenate was filtered through one layer of 20 µm Nitex membrane (Millipore #NY2004700), centrifuged at 3500 rpm for 3 min and the pellet were resuspended in nuclear freezing buffer (50 mM Tris at pH 8.0, 25% glycerol, 5 mM Mg(OAc)2, 0.1 mM EDTA, 5 mM DTT, 1× protease inhibitor cocktail (Roche #11697498001)). An aliquot of 50 µL was mixed with 50 µL of CountBright Absolute Counting Beads (ThermoFisher # C36950) and 3 µM DAPI and the nuclei were counted on an LSRFortessa cytometer. Based on the cytometer quantification, isolated nuclei were split into aliquots of ten million, snap frozen in liquid nitrogen and stored at -80 °C until use.

2.3.2 Nuclear staining for sorting of mesoderm / muscle populations by FACS

One day prior to the sci-ATAC-seq experiments, aliquots of 10 million nuclei obtained from wild-type, *bap* and *bin* collections were prepared for Fluorescence-Activated Cell Sorting (FACS) using an improved BiTS protocol as described previously (Reddington

et al., 2020). Nuclei were thawed and washed twice in PBT (PBS + 0.1% Triton-X) by centrifugation at 3500 rpm at 4 °C for 3 min, followed by overnight primary antibody staining at 4 °C in 400 µL 1X PBS supplemented with 5% BSA, 0.1% TritonX-100 and 1× protease inhibitor cocktail (Roche #11697498001). Rabbit primary antibodies anti-Mef2 and anti-Biniou (both from (Reddington et al., 2020); 1:1000 dilution) were used to mark the early mesoderm or myogenic mesoderm and to mark the visceral muscle respectively. Secondary antibody staining was performed by incubation with fluorescently labelled donkey anti-rabbit IgG-PE conjugate (Biolegend #406421; 1:200 dilution) for 1 hour at 4 °C in 400 µl 1X PBS supplemented with 5% BSA, 0.1% TritonX-100 and 1× protease inhibitor cocktail (Roche #11697498001).

2.4 sci-ATAC-seq

2.4.1 Generation of Tn5 transposomes for combinatorial indexing

For the standard protocol, pre-indexed Tn5 was obtained from Illumina. For the homemade protocol (described in Chapter 3), hyperactive Tn5 transposase was purified by the EMBL Protein Expression and Purification facility as previously described (Rossi et al., 2018) and stored at −20 °C in storage buffer (25 mM Tris pH 7.5, 800 mM NaCl,0.1 mM EDTA, 1 mM DTT, 50% glycerol) until use. Uniquely indexed oligonucleotides from (Cusanovich et al., 2018a) (Table 3) were annealed to common pMENTs oligos (95 °C 5 min, cooling to 65 °C (0.1 °C /sec), 65 °C 5 min, cooling to 4 °C (0.1 °C /sec)) to generate indexed transposons that were then loaded onto purified Tn5 by incubation at 23 °C with shaking at 350 rpm for 30 minutes. The loaded Tn5 transposomes were diluted 1:10 (final 0.02 mg/ml) in nuclease-free water and used immediately for tagmentation.

Table 3. Tn5 oligos sequence.

Oligo name	Oligo sequence (5'-3')
P5_i5_1_Universal_Connector_A_C15_ME	TCGTCGGCAGCGTCTCCACGCTATAGCCTGCGATCGAGGACGGCAGATGTGTATAAGAGACAG
P5_i5_2_Universal_Connector_A_C15_ME	TCGTCGGCAGCGTCTCCACGCATAGAGGCGCGATCGAGGACGGCAGATGTGTATAAGAGACAG
P5_i5_3_Universal_Connector_A_C15_ME	TCGTCGGCAGCGTCTCCACGCCCTATCCTGCGATCGAGGACGGCAGATGTGTATAAGAGACAG
P5_i5_4_Universal_Connector_A_C15_ME	TCGTCGGCAGCGTCTCCACGCGGCTCTGAGCGATCGAGGACGGCAGATGTGTATAAGAGACAG
P5_i5_5_Universal_Connector_A_C15_ME	TCGTCGGCAGCGTCTCCACGCAGGCGAAGGCGATCGAGGACGGCAGATGTGTATAAGAGACAG
P5_i5_6_Universal_Connector_A_C15_ME	TCGTCGGCAGCGTCTCCACGCTAATCTTAGCGATCGAGGACGGCAGATGTGTATAAGAGACAG
P5_i5_7_Universal_Connector_A_C15_ME	TCGTCGGCAGCGTCTCCACGCCAGGACGTGCGATCGAGGACGGCAGATGTGTATAAGAGACAG
P5_i5_8_Universal_Connector_A_C15_ME	TCGTCGGCAGCGTCTCCACGCGTACTGACGCGATCGAGGACGGCAGATGTGTATAAGAGACAG
P7_i7_1_Universal_Connector_B_D15_ME	GTCTCGTGGGCTCGGCTGTCCCTGTCCCGAGTAATCACCGTCTCCGCCTCAGATGTGTATAAGAGACAG
P7_i7_2_Universal_Connector_B_D15_ME	GTCTCGTGGGCTCGGCTGTCCCTGTCCTCTCCGGACACCGTCTCCGCCTCAGATGTGTATAAGAGACAG
P7_i7_3_Universal_Connector_B_D15_ME	GTCTCGTGGGCTCGGCTGTCCCTGTCCAATGAGCGCACCGTCTCCGCCTCAGATGTGTATAAGAGACAG
P7_i7_4_Universal_Connector_B_D15_ME	GTCTCGTGGGCTCGGCTGTCCCTGTCCGGAATCTCCACCGTCTCCGCCTCAGATGTGTATAAGAGACAG
P7_i7_5_Universal_Connector_B_D15_ME	GTCTCGTGGGCTCGGCTGTCCCTGTCCTTCTGAATCACCGTCTCCGCCTCAGATGTGTATAAGAGACAG
P7_i7_6_Universal_Connector_B_D15_ME	GTCTCGTGGGCTCGGCTGTCCCTGTCCACGAATTCCACCGTCTCCGCCTCAGATGTGTATAAGAGACAG
P7_i7_7_Universal_Connector_B_D15_ME	GTCTCGTGGGCTCGGCTGTCCCTGTCCAGCTTCAGCACCGTCTCCGCCTCAGATGTGTATAAGAGACAG
P7_i7_8_Universal_Connector_B_D15_ME	GTCTCGTGGGCTCGGCTGTCCCTGTCCGCGCATTACACCGTCTCCGCCTCAGATGTGTATAAGAGACAG
P7_i7_9_Universal_Connector_B_D15_ME	GTCTCGTGGGCTCGGCTGTCCCTGTCCCATAGCCGCACCGTCTCCGCCTCAGATGTGTATAAGAGACAG
P7_i7_10_Universal_Connector_B_D15_ME	GTCTCGTGGGCTCGGCTGTCCCTGTCCCATAGCCGCACCGTCTCCGCCTCAGATGTGTATAAGAGACAG
P7_i7_11_Universal_Connector_B_D15_ME	GTCTCGTGGGCTCGGCTGTCCCTGTCCGCGCGAGACACCGTCTCCGCCTCAGATGTGTATAAGAGACAG
P7_i7_12_Universal_Connector_B_D15_ME	GTCTCGTGGGCTCGGCTGTCCCTGTCCCTATCGCTCACCGTCTCCGCCTCAGATGTGTATAAGAGACAG
pMENTS	5Phos/CTGTCTCTTATACACATCT

2.4.2 Generation of sci-ATAC-seq libraries

Generation of sci-ATAC-seq libraries was performed largely as previously described (Cusanovich et al., 2018a), with some modifications. Nuclei were washed twice by pelleting and resuspending in 1 mL 1X PBS supplemented with 0.1% TritonX-100 and

1× protease inhibitor cocktail (Roche #11697498001). Nuclei were stained with 3 µM DAPI and 2,500 DAPI+ (and Mef2+ or Biniou+ for mesoderm / visceral muscle sorting) nuclei were sorted into each well of a 96-well plate containing 5 µL of Omni-ATAC buffer (Corces et al., 2017) supplemented with 1× protease inhibitor cocktail (Roche #11697498001) and 12 µL of TD buffer (Illumina) in each well. Tagmentation was performed by adding 2 µl of each of the 96 custom and uniquely indexed Tn5 transposomes (Illumina Tn5 for the standard protocol, EMBL Tn5 for the homemade protocol) and incubating at 55 °C for 2 hours. After tagmentation, nuclei were pooled in a 15 mL falcon tube, stained with 3 µM DAPI and 25 DAPI+ nuclei were sorted into each well of 96-well plates. Following reverse-crosslinking, tagmented DNA was PCR amplified by adding 5 µL of 5 µM forward and reverse indexed primers (Table 4) (Cusanovich et al., 2018a), 7.5 µl of NPM polymerase master mix (Illumina) and BSA (2X final concentration; NEB) to each well and by running the following cycling conditions: 72 °C 5 min, 98 °C 30 s; 98 °C 10 s, 63 °C 30 s, 19–20 cycles; 72 °C 1 min, hold at 10 °C. For PCR amplification with the homemade protocol (as described in Chapter 3), the NPM master mix was replaced with 7.5 µL KAPA HiFi DNA Polymerase ReadyMix (Roche #7958935001). The optimal number of cycles for each library was determined beforehand by monitoring amplification on a qPCR machine for a set of test wells.

Table 4. Indexed PCR oligos sequence.

Oligo name	Oligo sequence (5'-3')
P5_1_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACCTCCATCGAGTCGTCGGCAGCGTC
P5_2_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACTTGGTAGTCGTCGTCGGCAGCGTC
P5_3_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACGGCCGTCAACTCGTCGGCAGCGTC
P5_4_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACCCTAGACGAGTCGTCGGCAGCGTC
P5_5_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACTCGTTAGAGCTCGTCGGCAGCGTC
P5_6_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACCGTTCTATCATCGTCGGCAGCGTC
P5_7_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACCGGAATCTAATCGTCGGCAGCGTC
P5_8_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACATGACTGATCTCGTCGGCAGCGTC
P5_9_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACTCAATATCGATCGTCGGCAGCGTC
P5_10_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACGTAGACCTGGTCGTCGGCAGCGTC
P5_11_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACTTATGACCAATCGTCGGCAGCGTC
P5_12_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACTTGGTCCGTTTCGTCGGCAGCGTC
P5_13_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACGGTACGTTAATCGTCGGCAGCGTC
P5_14_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACCAATGAGTCCTCGTCGGCAGCGTC
P5_15_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACGATGCAGTTCTCGTCGGCAGCGTC
P5_16_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACCCATCGTTCCTCGTCGGCAGCGTC

P5_17_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACTTGAGAGAGTTCGTCGGCAGCGTC
P5_18_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACACTGAGCGACTCGTCGGCAGCGTC
P5_19_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACTGAGGAATCATCGTCGGCAGCGTC
P5_20_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACCCTCCGACGGTCGTCGGCAGCGTC
P5_21_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACCATTGACGCTTCGTCGGCAGCGTC
P5_22_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACTCGTCCTTCGTCGTCGGCAGCGTC
P5_23_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACTGATACTCAATCGTCGGCAGCGTC
P5_24_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACTTCTACCTCATCGTCGGCAGCGTC
P5_25_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACTCGTCGGAACTCGTCGGCAGCGTC
P5_26_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACATCGAGATGATCGTCGGCAGCGTC
P5_27_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACTAGACTAGTCTCGTCGGCAGCGTC
P5_28_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACGTCGAAGCAGTCGTCGGCAGCGTC
P5_29_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACAGGCGCTAGGTCGTCGGCAGCGTC
P5_30_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACAGATGCAACTTCGTCGGCAGCGTC
P5_31_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACAAGCCTACGATCGTCGGCAGCGTC
P5_32_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACGTAGGCAATTTCGTCGGCAGCGTC
P5_33_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACGGAGGCGGCGTCGTCGGCAGCGTC
P5_34_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACCCAGTACTTGTCGTCGGCAGCGTC
P5_35_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACGGTCTCGCCGTCGTCGGCAGCGTC
P5_36_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACGGCGGAGGTCTCGTCGGCAGCGTC
P5_37_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACTAGTTCTAGATCGTCGGCAGCGTC
P5_38_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACTTGGAGTTAGTCGTCGGCAGCGTC
P5_39_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACAGATCTTGGTTCGTCGGCAGCGTC
P5_40_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACGTAATGATCGTCGTCGGCAGCGTC
P5_41_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACCAGAGAGGTCTCGTCGGCAGCGTC
P5_42_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACTTAATTAGCCTCGTCGGCAGCGTC
P5_43_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACCTCTAACTCGTCGTCGGCAGCGTC
P5_44_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACTACGATCATCTCGTCGGCAGCGTC
P5_45_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACAGGCGAGAGCTCGTCGGCAGCGTC
P5_46_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACTCAAGATAGTTCGTCGGCAGCGTC
P5_47_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACTAATTGACCTTCGTCGGCAGCGTC
P5_48_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACCAGCCGGCTTTCGTCGGCAGCGTC
P5_49_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACAGAACCGGAGTCGTCGGCAGCGTC
P5_50_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACGAGATGCATGTCGTCGGCAGCGTC
P5_51_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACGATTACCGGATCGTCGGCAGCGTC
P5_52_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACTCGTAACGGTTCGTCGGCAGCGTC
P5_53_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACTGGCGACGGATCGTCGGCAGCGTC
P5_54_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACAGTCATAGCCTCGTCGGCAGCGTC
P5_55_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACGTCAAGTCCATCGTCGGCAGCGTC
P5_56_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACATTCGGAAGTTCGTCGGCAGCGTC
P5_57_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACGTCGGTAGTTTCGTCGGCAGCGTC
P5_58_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACAGGACGGACGTCGTCGGCAGCGTC
P5_59_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACCTCCTGGACCTCGTCGGCAGCGTC
P5_60_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACTAGCCTCGTTTCGTCGGCAGCGTC
P5_61_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACGGTTGAACGTTCGTCGGCAGCGTC

P5_62_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACAGGTCCTCGTTCGTCGGCAGCGTC
P5_63_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACGGAAGTTATATCGTCGGCAGCGTC
P5_64_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACTGGTAATCCTTCGTCGGCAGCGTC
P5_65_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACAAGCTAGGTTTCGTCGGCAGCGTC
P5_66_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACTCCGCGGACTTCGTCGGCAGCGTC
P5_67_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACTGCGGATAGTTCGTCGGCAGCGTC
P5_68_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACTGGCAGCTCGTCGTCGGCAGCGTC
P5_69_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACTGCTACGGTCTCGTCGGCAGCGTC
P5_70_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACGCGCAATGACTCGTCGGCAGCGTC
P5_71_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACCTTAATCTTGTCGTCGGCAGCGTC
P5_72_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACGGAGTTGCGTTCGTCGGCAGCGTC
P5_73_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACACTCGTATCATCGTCGGCAGCGTC
P5_74_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACGGTAATAATGTCGTCGGCAGCGTC
P5_75_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACTCCTTATAGATCGTCGGCAGCGTC
P5_76_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACCCGACTCCAATCGTCGGCAGCGTC
P5_77_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACGCCAAGCTTGTCGTCGGCAGCGTC
P5_78_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACCATATCCTATTCGTCGGCAGCGTC
P5_79_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACACCTACGCCATCGTCGGCAGCGTC
P5_80_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACGGAATTCAGTTCGTCGGCAGCGTC
P5_81_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACTGGCGTAGAATCGTCGGCAGCGTC
P5_82_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACATTGCGGCCATCGTCGGCAGCGTC
P5_83_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACTTCAGCTTGGTCGTCGGCAGCGTC
P5_84_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACCCATCTGGCATCGTCGGCAGCGTC
P5_85_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACCTTATAAGTTTCGTCGGCAGCGTC
P5_86_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACGATTAGATGATCGTCGGCAGCGTC
P5_87_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACTATAGGATCTTCGTCGGCAGCGTC
P5_88_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACAGCTTATAGGTCGTCGGCAGCGTC
P5_89_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACGTCTGCAATCTCGTCGGCAGCGTC
P5_90_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACCGCCTCTTATTCGTCGGCAGCGTC
P5_91_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACGTTGGATCTTTCGTCGGCAGCGTC
P5_92_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACGCGATTGCAGTCGTCGGCAGCGTC
P5_93_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACTGCCAGTTGCTCGTCGGCAGCGTC
P5_94_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACCTTAGGTATCTCGTCGGCAGCGTC
P5_95_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACGAGACCTACCTCGTCGGCAGCGTC
P5_96_PCR_Primer	AATGATACGGCGACCACCGAGATCTACACATTGACCGAGTCGTCGGCAGCGTC
P7_1_PCR_Primer	CAAGCAGAAGACGGCATACGAGATCCGAATCCGAGTCTCGTGGGCTCGG
P7_2_PCR_Primer	CAAGCAGAAGACGGCATACGAGATCCGCAGCCGCGTCTCGTGGGCTCGG
P7_3_PCR_Primer	CAAGCAGAAGACGGCATACGAGATAACGTAATCTGTCTCGTGGGCTCGG
P7_4_PCR_Primer	CAAGCAGAAGACGGCATACGAGATACCTAGTTAGGTCTCGTGGGCTCGG
P7_5_PCR_Primer	CAAGCAGAAGACGGCATACGAGATGGTCGCTATGGTCTCGTGGGCTCGG
P7_6_PCR_Primer	CAAGCAGAAGACGGCATACGAGATCTCTTAGCGGGTCTCGTGGGCTCGG
P7_7_PCR_Primer	CAAGCAGAAGACGGCATACGAGATTTCGTTCCATGTCTCGTGGGCTCGG
P7_8_PCR_Primer	CAAGCAGAAGACGGCATACGAGATAACGGAACGCGTCTCGTGGGCTCGG
P7_9_PCR_Primer	CAAGCAGAAGACGGCATACGAGATTTCGATAACCGTCTCGTGGGCTCGG
P7_10_PCR_Primer	CAAGCAGAAGACGGCATACGAGATCATACGATGCGTCTCGTGGGCTCGG

P7_11_PCR_Primer	CAAGCAGAAGACGGCATACGAGATTTATCGTATTGTCTCGTGGGCTCGG
P7_12_PCR_Primer	CAAGCAGAAGACGGCATACGAGATGTCGACGGAAGTCTCGTGGGCTCGG
P7_13_PCR_Primer	CAAGCAGAAGACGGCATACGAGATATAAGCCGGAGTCTCGTGGGCTCGG
P7_14_PCR_Primer	CAAGCAGAAGACGGCATACGAGATTGCGCCTGGTGTCTCGTGGGCTCGG
P7_15_PCR_Primer	CAAGCAGAAGACGGCATACGAGATATTCTCCTCTGTCTCGTGGGCTCGG
P7_16_PCR_Primer	CAAGCAGAAGACGGCATACGAGATATAGGAGTACGTCTCGTGGGCTCGG
P7_17_PCR_Primer	CAAGCAGAAGACGGCATACGAGATATCCGTTAGCGTCTCGTGGGCTCGG
P7_18_PCR_Primer	CAAGCAGAAGACGGCATACGAGATTGATTCAACTGTCTCGTGGGCTCGG
P7_19_PCR_Primer	CAAGCAGAAGACGGCATACGAGATTACCTAATCAGTCTCGTGGGCTCGG
P7_20_PCR_Primer	CAAGCAGAAGACGGCATACGAGATGATGCTACGAGTCTCGTGGGCTCGG
P7_21_PCR_Primer	CAAGCAGAAGACGGCATACGAGATAACCTCAAGAGTCTCGTGGGCTCGG
P7_22_PCR_Primer	CAAGCAGAAGACGGCATACGAGATAAGCTGACCTGTCTCGTGGGCTCGG
P7_23_PCR_Primer	CAAGCAGAAGACGGCATACGAGATAAGTCTAATAGTCTCGTGGGCTCGG
P7_24_PCR_Primer	CAAGCAGAAGACGGCATACGAGATACTAATTGAGGTCTCGTGGGCTCGG
P7_25_PCR_Primer	CAAGCAGAAGACGGCATACGAGATCCGGCGGCGAGTCTCGTGGGCTCGG
P7_26_PCR_Primer	CAAGCAGAAGACGGCATACGAGATAATCATACGGGTCTCGTGGGCTCGG
P7_27_PCR_Primer	CAAGCAGAAGACGGCATACGAGATTCTGCGCGTTGTCTCGTGGGCTCGG
P7_28_PCR_Primer	CAAGCAGAAGACGGCATACGAGATCTACGACGAGGTCTCGTGGGCTCGG
P7_29_PCR_Primer	CAAGCAGAAGACGGCATACGAGATTCGCAATTAGGTCTCGTGGGCTCGG
P7_30_PCR_Primer	CAAGCAGAAGACGGCATACGAGATTATGGCCGCGGTCTCGTGGGCTCGG
P7_31_PCR_Primer	CAAGCAGAAGACGGCATACGAGATAAGTAATATTGTCTCGTGGGCTCGG
P7_32_PCR_Primer	CAAGCAGAAGACGGCATACGAGATATCTGCCAATGTCTCGTGGGCTCGG
P7_33_PCR_Primer	CAAGCAGAAGACGGCATACGAGATCAGGCGCCATGTCTCGTGGGCTCGG
P7_34_PCR_Primer	CAAGCAGAAGACGGCATACGAGATGAGTCCTTATGTCTCGTGGGCTCGG
P7_35_PCR_Primer	CAAGCAGAAGACGGCATACGAGATCGGCTTACTAGTCTCGTGGGCTCGG
P7_36_PCR_Primer	CAAGCAGAAGACGGCATACGAGATCTTGCATAATGTCTCGTGGGCTCGG
P7_37_PCR_Primer	CAAGCAGAAGACGGCATACGAGATGGCTTGCCAAGTCTCGTGGGCTCGG
P7_38_PCR_Primer	CAAGCAGAAGACGGCATACGAGATCGCCAATCAAGTCTCGTGGGCTCGG
P7_39_PCR_Primer	CAAGCAGAAGACGGCATACGAGATGCTCATATGCGTCTCGTGGGCTCGG
P7_40_PCR_Primer	CAAGCAGAAGACGGCATACGAGATAGTCGAGTTCGTCTCGTGGGCTCGG
P7_41_PCR_Primer	CAAGCAGAAGACGGCATACGAGATGGCTGGCTAGGTCTCGTGGGCTCGG
P7_42_PCR_Primer	CAAGCAGAAGACGGCATACGAGATAGAGGTCGCAGTCTCGTGGGCTCGG
P7_43_PCR_Primer	CAAGCAGAAGACGGCATACGAGATAGCTAAGAATGTCTCGTGGGCTCGG
P7_44_PCR_Primer	CAAGCAGAAGACGGCATACGAGATATCGTATCAAGTCTCGTGGGCTCGG
P7_45_PCR_Primer	CAAGCAGAAGACGGCATACGAGATAACTATTATAGTCTCGTGGGCTCGG
P7_46_PCR_Primer	CAAGCAGAAGACGGCATACGAGATCCTACGGCAAGTCTCGTGGGCTCGG
P7_47_PCR_Primer	CAAGCAGAAGACGGCATACGAGATGATATGGTCTGTCTCGTGGGCTCGG
P7_48_PCR_Primer	CAAGCAGAAGACGGCATACGAGATTCCTTACCAAGTCTCGTGGGCTCGG
P7_49_PCR_Primer	CAAGCAGAAGACGGCATACGAGATCCGCTAGCTGGTCTCGTGGGCTCGG
P7_50_PCR_Primer	CAAGCAGAAGACGGCATACGAGATCAAGGCTTAGGTCTCGTGGGCTCGG
P7_51_PCR_Primer	CAAGCAGAAGACGGCATACGAGATAGCGGTAACGGTCTCGTGGGCTCGG
P7_52_PCR_Primer	CAAGCAGAAGACGGCATACGAGATTGGTCCAGTCGTCTCGTGGGCTCGG
P7_53_PCR_Primer	CAAGCAGAAGACGGCATACGAGATACGGTCTTGCGTCTCGTGGGCTCGG

P7_54_PCR_Primer	CAAGCAGAAGACGGCATACGAGATAGGAGATTGAGTCTCGTGGGCTCGG
P7_55_PCR_Primer	CAAGCAGAAGACGGCATACGAGATGTCGAGGTATGTCTCGTGGGCTCGG
P7_56_PCR_Primer	CAAGCAGAAGACGGCATACGAGATAACGCCTCTAGTCTCGTGGGCTCGG
P7_57_PCR_Primer	CAAGCAGAAGACGGCATACGAGATAAGTTACCTAGTCTCGTGGGCTCGG
P7_58_PCR_Primer	CAAGCAGAAGACGGCATACGAGATAATATTCGAAGTCTCGTGGGCTCGG
P7_59_PCR_Primer	CAAGCAGAAGACGGCATACGAGATTAGTCGTCCAGTCTCGTGGGCTCGG
P7_60_PCR_Primer	CAAGCAGAAGACGGCATACGAGATTGCAGCCTACGTCTCGTGGGCTCGG
P7_61_PCR_Primer	CAAGCAGAAGACGGCATACGAGATCTTATCCTACGTCTCGTGGGCTCGG
P7_62_PCR_Primer	CAAGCAGAAGACGGCATACGAGATGCGCTCGACGGTCTCGTGGGCTCGG
P7_63_PCR_Primer	CAAGCAGAAGACGGCATACGAGATAATGAATAGTGTCTCGTGGGCTCGG
P7_64_PCR_Primer	CAAGCAGAAGACGGCATACGAGATATCTAAGCAAGTCTCGTGGGCTCGG
P7_65_PCR_Primer	CAAGCAGAAGACGGCATACGAGATGCTCCATTCGGTCTCGTGGGCTCGG
P7_66_PCR_Primer	CAAGCAGAAGACGGCATACGAGATGGCTATATAGGTCTCGTGGGCTCGG
P7_67_PCR_Primer	CAAGCAGAAGACGGCATACGAGATTTATTAGTAGGTCTCGTGGGCTCGG
P7_68_PCR_Primer	CAAGCAGAAGACGGCATACGAGATACGGCAACCAGTCTCGTGGGCTCGG
P7_69_PCR_Primer	CAAGCAGAAGACGGCATACGAGATCGGCAGAGGAGTCTCGTGGGCTCGG
P7_70_PCR_Primer	CAAGCAGAAGACGGCATACGAGATTTCAAGAATCGTCTCGTGGGCTCGG
P7_71_PCR_Primer	CAAGCAGAAGACGGCATACGAGATTAGCTGCTACGTCTCGTGGGCTCGG
P7_72_PCR_Primer	CAAGCAGAAGACGGCATACGAGATGGAGCTGAGGGTCTCGTGGGCTCGG
P7_73_PCR_Primer	CAAGCAGAAGACGGCATACGAGATTGAGCTACTTGTCTCGTGGGCTCGG
P7_74_PCR_Primer	CAAGCAGAAGACGGCATACGAGATTCCAGCAATAGTCTCGTGGGCTCGG
P7_75_PCR_Primer	CAAGCAGAAGACGGCATACGAGATCCGTATCTGGGTCTCGTGGGCTCGG
P7_76_PCR_Primer	CAAGCAGAAGACGGCATACGAGATCGAATTCGTTGTCTCGTGGGCTCGG
P7_77_PCR_Primer	CAAGCAGAAGACGGCATACGAGATACGATAAGCGGTCTCGTGGGCTCGG
P7_78_PCR_Primer	CAAGCAGAAGACGGCATACGAGATTCGCGTACTTGTCTCGTGGGCTCGG
P7_79_PCR_Primer	CAAGCAGAAGACGGCATACGAGATTGCGAAGATCGTCTCGTGGGCTCGG
P7_80_PCR_Primer	CAAGCAGAAGACGGCATACGAGATCAGGCTAAGAGTCTCGTGGGCTCGG
P7_81_PCR_Primer	CAAGCAGAAGACGGCATACGAGATGCCTCAATAAGTCTCGTGGGCTCGG
P7_82_PCR_Primer	CAAGCAGAAGACGGCATACGAGATATGCTCGCAAGTCTCGTGGGCTCGG
P7_83_PCR_Primer	CAAGCAGAAGACGGCATACGAGATCTCTTCAAGCGTCTCGTGGGCTCGG
P7_84_PCR_Primer	CAAGCAGAAGACGGCATACGAGATGCAGCGGACTGTCTCGTGGGCTCGG
P7_85_PCR_Primer	CAAGCAGAAGACGGCATACGAGATTCAGGACTTAGTCTCGTGGGCTCGG
P7_86_PCR_Primer	CAAGCAGAAGACGGCATACGAGATCATGAGAACTGTCTCGTGGGCTCGG
P7_87_PCR_Primer	CAAGCAGAAGACGGCATACGAGATCCTTAGTCTGGTCTCGTGGGCTCGG
P7_88_PCR_Primer	CAAGCAGAAGACGGCATACGAGATCAGCGATAGAGTCTCGTGGGCTCGG
P7_89_PCR_Primer	CAAGCAGAAGACGGCATACGAGATACCATAGCGCGTCTCGTGGGCTCGG
P7_90_PCR_Primer	CAAGCAGAAGACGGCATACGAGATAATAATAATGGTCTCGTGGGCTCGG
P7_91_PCR_Primer	CAAGCAGAAGACGGCATACGAGATAACTACGGCTGTCTCGTGGGCTCGG
P7_92_PCR_Primer	CAAGCAGAAGACGGCATACGAGATCGCAATATCAGTCTCGTGGGCTCGG
P7_93_PCR_Primer	CAAGCAGAAGACGGCATACGAGATTTAACGCCGTGTCTCGTGGGCTCGG
P7_94_PCR_Primer	CAAGCAGAAGACGGCATACGAGATGGAGTAAGCCGTCTCGTGGGCTCGG
P7_95_PCR_Primer	CAAGCAGAAGACGGCATACGAGATATGAACGCGCGTCTCGTGGGCTCGG
P7_96_PCR_Primer	CAAGCAGAAGACGGCATACGAGATCATCGCGCTCGTCTCGTGGGCTCGG

2.4.3 Library clean-up and sequencing

After PCR amplification, library clean-up was performed with DNA Clean & Concentrator-5 Kit (Zymo #D4014), following the manufacturer's instructions. DNA concentration was quantified with Qubit dsDNA HS Assay Kit (ThermoFisher # Q32851) and 1 µL of each library was loaded on Novex 6% TBE gels (Invitrogen # EC62652BOX) to visualize the DNA size distribution. Libraries were sequenced on an Illumina NextSeq 500 sequencer High Capacity 150 PE kit (Illumina) using custom primers (Table 5) and a custom sequencing recipe previously described in (Amini et al., 2014). All sequencing was performed by staff at the EMBL Genecore Facility.

Table 5. Sequencing primers sequence.

Oligo name	Oligo sequence (5'-3')
Read 1 Sequencing Primer	GCGATCGAGGACGGCAGATGTGTATAAGAGACAG
Read 2 Sequencing Primer	CACCGTCTCCGCCTCAGATGTGTATAAGAGACAG
Index 1 Sequencing Primer	CTGTCTCTTATACACATCTGAGGCGGAGACGGTG
Index 2 Sequencing Primer	CTGTCTCTTATACACATCTGCCGTCCTCGATCGC

2.5 Raw sci-ATAC-seq data processing and cell assignment

2.5.1 Raw sequencing data processing, mapping and duplicate removal

Processing of raw sequencing data was performed with the scripts and pipeline provided in (Cusanovich et al., 2018a). Conversion of BCL files into fastq files was done using bcl2fastq v.2.16 (Illumina). Read barcodes presenting sequencing or PCR amplification errors were first corrected to their presumptive match (Levenshtein distance < 3 and distance to next best match > 2); all other barcodes were classified as ambiguous or unknown. Reads were trimmed with Trimmomatic v0.32 (Bolger et al., 2014) and mapped to the dm6 reference genome using Bowtie2 v2.3.4.1 (Langmead and Salzberg, 2012) (with options -X 2000 -3 1). Reads that did not map to autosomes or sex chromosomes with a minimum mapping quality of 10 and reads having ambiguous or unknown barcodes, were discarded. PCR duplicates were

removed with the python script ‘sc_atac_library_deconvoluter.py’ from (Cusanovich et al., 2018a).

2.5.2 Barcode-cell assignment and quality control

After removal of PCR duplicates, I proceeded to identify cells and exclude low quality cells by applying three stringent filters: (1) barcodes were classified as genuine cells if there was no more than 5% uncertainty that they belonged to the higher read depth component of a Gaussian mixture model fitted to the distribution of read counts per barcode (as in (Cusanovich et al., 2018a)). (2) As the insert size distribution was evident even in single cells (Figure 6e, Chapter 3), I quantified nucleosomal banding per-cell using a fast Fourier transform-based metric with the scripts provided in (Cusanovich et al., 2018b) and retained cells with clear nucleosomal banding. (3) I applied a third filter to remove cells missing a clear sub-nucleosomal peak, based on per-cell quantification of sub-nucleosomal fragments.

2.6 Count matrix generation, clustering and visualization

2.6.1 Peak calling and generation of bigwig tracks

Deconvoluted BAM files were generated for each described sample / condition with the script ‘sc_atac_library_deconvoluter.py’ from (Cusanovich et al., 2018a) and were then converted to bigwig files with deepTools ‘bamCoverage’ v2.5.1 (normalization option: --normalizeUsingRPKM). Deconvoluted BAM files were also used for peak calling, which in all instances was performed with MACS2 v2.2.7.1 (Zhang et al., 2008), using the ‘macs2 callpeak’ command with parameters: ‘--nomodel --keep-dup all --extsize 200 --shift -100’. Peak summits were resized to 300 bp and merged with BEDTools v2.27.1 (Quinlan and Hall, 2010) to generate master lists of peaks for each dataset, as described in the following sections.

2.6.2 LDA, clustering and UMAP visualization

Accessibility matrices of peaks (rows) by cells (columns) for each described sample / condition were generated with script 'sc_atac_window_counter.py' from (Cusanovich et al., 2018a) and subsequently used as input for Latent Dirichlet Allocation (LDA) with cisTopic function 'runModels' (cisTopic v0.2.2 (Bravo González-Blas et al., 2019), R package). Sex chromosomes were removed prior to LDA to avoid sex bias in the clustering, as recommended in (Cusanovich et al., 2018a). The resulting topic by cell matrix was fed into Seurat (v3.2.2, R package) as dimensionality reduction for computing UMAP plots (function 'RunUMAP') and clustering (function 'FindClusters').

2.6.3 Gene activity matrix generation

Gene activity scores were calculated with script 'sc_atac_window_counter.py' from (Cusanovich et al., 2018a) by computing accessibility over the whole gene body plus 500 bp upstream, for genes in the R Bionconductor package 'TxDb.Dmelanogaster.UCSC.dm6.ensGene_3.4.4'. The resulting gene by cell accessibility matrices were imported in Seurat for downstream analysis.

2.7 Cell-type annotation

ATAC-seq peaks and genes were tested for differential accessibility in each cluster by logistic regression using Seurat v3.2.2 function 'FindAllMarkers', with the total counts per cell given as a latent variable. Features with a positive log fold-change and a Bonferroni adjusted P-value below 0.05 were considered to be markers of a given cluster, while the non-significant features were retained as background. The cluster markers and the background features were matched to activity terms of characterized *in vivo* enhancer activity (Bonn et al., 2012a; Kvon et al., 2014; Rivera et al., 2019) or BDGP gene expression (Tomancak et al., 2002), and each activity term in a given cluster was tested for over-representation against the background with a Fisher's one-tailed test. As many of the activity terms are highly overlapping, the Fisher's test p-

values were not formally corrected for multiple comparison, and instead to assign cell types I focused on large and consistent enrichments of similar activity terms.

2.8 Analysis of the wild-type mesoderm time course

2.8.1 Clustering the wild-type mesoderm time course dataset

After cell filtering based on the QC metrics described above, the two sci-ATAC-seq replicates were merged giving a combined data set of 24,032 wild-type cells. To maximise the resolution of the mesoderm/muscle chromatin accessibility landscape over the whole time course, I performed two rounds of clustering: first each time point was clustered individually with Seurat v3.2.2 function 'FindClusters' based on accessibility quantified on a merged set of 42,076 peaks that were called separately for each time point. This process allowed the identification of 61 Seurat clusters, from which I excluded low quality and suspected collision clusters (12 clusters comprising 2,725 cells) based on the read depth and sex ratio metrics, as described in (Cusanovich et al., 2018a). Accessibility was quantified again at 50,261 merged peaks identified separately for each cluster and the resulting count matrix was used for clustering the full time course. Batch correction was not necessary, as I did not observe a clustering bias for the two replicates.

2.8.2 Calculation of transcription factor deviation scores

To investigate the relationship between accessibility changes and transcription factor occupancy, I retrieved 16 ChIP datasets of mesodermal factors from our lab (Cunha et al., 2010; Jakobsen et al., 2007; Junion et al., 2012; Zinzen et al., 2009) and 280 ChIP datasets on diverse factors from the modERN database (Kudron et al., 2018). For analysis with the modERN data, I followed the previously reported strategy (Reddington et al., 2020) to exclude ATAC-seq peaks occupied by any TF with ubiquitous expression, which resulted in 30,318 peaks being retained for analysis from the original list of 50,261 peaks. Deviations in accessibility were calculated with chromVAR v1.10.0 (Schep et al., 2017).

2.8.3 Reconstruction of lineage trajectories and pseudotime analysis

To order cells in pseudotime, I identified trajectories for the myogenic lineages and aligned single cells along the trajectories following the approach outlined in (Satpathy et al., 2019). I used the function 'alignTrajectory' provided in (Satpathy et al., 2019) to construct trajectories in the UMAP subspace for the somatic (clusters 5, 0, 2, 3, 6), cardiac (clusters 5, 0, 15) and visceral lineages (clusters 5, 13, 8, 7) and to calculate pseudotime along the aligned cells. I used the functions 'getTrajectory' and 'plotTrajectoryHeatmap' from ArchR v1.0 (Granja et al., 2021) to reconstruct feature trends across pseudotime and plot the peaks and genes heatmaps. The binarized accessibility matrix was used as input for peaks and the log-normalized gene activity matrix for genes. For heatmap visualization, I selected the top 10% and 20% most variable peaks and genes across pseudotime.

2.8.4 Identification of DA peaks and genes for muscle subpopulations

Logistic regression was performed to identify differentially accessible (DA) ATAC peaks and genes among muscle subpopulations using Seurat v3.2.2 function 'FindAllMarkers' (slot = counts, test.use = LR, logfc.threshold = 0, min.pct = 0.1, latent.vars = total counts in features per cell). Features with Bonferroni adjusted p-value < 0.001 and log2 fold-change > 0.5 were considered differentially accessible. The log2 fold-change was calculated using counts matrices scaled by the total counts in features per cell, in order to correct for potential coverage differences among clusters, and using a pseudocount of 10^{-6}. To generate heatmaps, the scaled accessibility was averaged per cluster with Seurat function 'AverageExpression'.

2.9 Single-nucleus *de-novo* genotyping

2.9.1 Defining a set of discriminatory variants for genotyping

A set of discriminatory variants, separating the Vasa-Cas9 and the Balancer chromosomes, was obtained using GATK version 4.1.0 (McKenna et al., 2010). In

brief, joint genotyping and variant calling was performed with GATK HaplotypeCaller (-G StandardAnnotation --min-base-quality-score=20), using the gDNA reads of the Balancer and the Vasa-Cas9 lines. The resulting variants were filtered with two sets of cutoffs to obtain a lenient and a stringent set of variants. Filters for stringent set: MQ > 58, MQRankSum > -2,5; MQRankSum < 2,5; QD > 20; SOR < 1.5; FS < 10; ReadPosRankSum > -4; ReadPosRankSum < 4. Filters for lenient set: MQ > 40; MQRankSum > -12,5; MQRankSum < 12,5; QD > 2; SOR < 3; FS < 60; ReadPosRankSum > -8; ReadPosRankSum < 8. This process yielded 104,913 single nucleotide variants (SNVs) separating the Vasa-Cas9 and Balancer chromosomes when applying the stringent filters and 465,110 when applying the lenient filters. This whole procedure was entirely performed by Dr. Mattia Forneris.

2.9.2 Genotype assignment

Genotypes were assigned using Vireo v0.5.6 (Huang et al., 2019) based on the lenient set of discriminatory SNVs identified between the Balancer chromosomes and the common isogenic Vasa-Cas9 genetic background of each mutant. This entire procedure was entirely performed by Tobias Heinen. The barcodes that passed the QC steps described above were further filtered for having been assigned a genotype, while unassigned nuclei were excluded from downstream analysis.

2.10 Analysis of TF mutant datasets

2.10.1 Clustering of the *Mef2* mutant dataset

After cell QC and genotype assignment, chromatin accessibility for *Mef2* nuclei was quantified at the 53,133 peaks previously identified in the whole-embryo sci-ATAC-seq dataset (Cusanovich et al., 2018a) and the mesoderm / muscle cells were identified by clustering as described above (5 clusters comprising 2,567 cells). Peaks were called as described above for each mesoderm / muscle cluster and merged with peaks called on each genotype, resulting in 54,609 merged peaks. This peak set was used to quantify accessibility in the 2,567 muscle cells and 739 additional muscle cells that

were identified by clustering the *Mef2* hand-sorted sample, totally 3,306 muscle cells. I performed one round of clustering followed by cell-type annotation. In the final clustering, I removed an unrelated neuronal population (78 cells) and a group of 216 interspersed cells that showed poor clustering based on silhouette analysis (Rousseeuw, 1987), a common method to evaluate cluster cohesion and separation (all clusters had an average silhouette width above zero (mean = 0.13) except the removed cells, which had a negative average silhouette width (mean = -0.09), as calculated with R function 'silhouette' from package 'cluster' v2.1.2). For co-clustering with the wild-type time course, chromatin accessibility was quantified at 66,105 peaks obtained from merging wild-type peaks (50,261) with the lists of *Mef2* genotype-cluster peaks defined above. Topic modelling was applied to the count matrix with cisTopic v0.2.2 (Bravo González-Blas et al., 2019) and the topic by cell matrix was batch corrected with Harmony v1.0 (Korsunsky et al., 2019) (theta = 0; wild-type cells were provided as reference with option 'reference_values') prior to UMAP visualization.

2.10.2 Clustering of *tinman*, *bagpipe* and *biniou* mutant datasets

After cell quality control and genotype assignment, chromatin accessibility for *bap* and *bin* mutant cells was initially quantified at the 50,261 peaks previously defined for the wild-type time course, and the resulting count matrices were clustered. Peaks were then called as described above for each cell cluster and merged with peaks called on each genotype, resulting in 57,878 and 55,490 merged peaks for *bap* and *bin* respectively. Similar to *Mef2* processing, chromatin accessibility for *tin* was quantified at 53,133 peaks previously identified in a whole-embryo sci-ATAC-seq dataset (Cusanovich et al., 2018a) and the mesoderm / muscle cells were identified by clustering (6 clusters comprising 6,786 cells). Peaks were called for each mesoderm / muscle cluster and merged with peaks called on each genotype, resulting in 60,104 merged peaks. For co-clustering with the wild-type time course, chromatin accessibility was quantified at 63,842 peaks obtained from merging wild-type peaks (50,261) with the lists of genotype-cluster peaks from each mutant of the three mutant datasets (*tin*, *bap*, *bin*). Topic modelling was applied to the count matrix using cisTopic v0.2.2 (Bravo González-Blas et al., 2019) and the topic-by-cell matrix was batch corrected with Harmony v1.0 (Korsunsky et al., 2019) (theta = 0; wild-type cells were provided as reference with option 'reference_values') prior to UMAP visualization and clustering.

Cell-type labels were very homogenous within each cluster, with an average of 80% of cells having the same label, therefore I assigned the most frequent label as the cell-type annotation for a given cluster. Significant imbalances in the proportion of homozygous mutant nuclei obtained in each cluster were identified using a Fisher's exact test against the observed overall proportion (Expected: 25%; Observed: 15%, 19%, 17% for *tin*, *bap* and *bin* respectively).

2.10.3 Differential ATAC peak analysis in *Mef2* mutant embryos

Logistic regression was performed to identify differentially accessible (DA) sites in Mutant1 cluster against the somatic cluster using Seurat v3.2.2 function 'FindMarkers' (test.use = LR, logfc.threshold = 0, min.pct = 0.1, latent.vars = total counts in peaks per cell). Out of 8,725 sites tested, 408 had significant differential accessibility (Bonferroni adjusted p-value < 0.05, log2 fold-change > +/- 0.5). To correct for potential coverage differences among clusters, the counts were scaled by the total counts in peaks per cell, prior to the log2 fold-change calculation. To generate heatmaps, accessibility was averaged per cluster with Seurat function 'AverageExpression'. Sites residing within 1 kb (+/- 500 bp) centered on a gene TSS were defined as gene-proximal, sites outside this region were considered gene-distal, and putative enhancers. BEDTools v2.27.1 (Quinlan and Hall, 2010) was used to find overlaps between sites and several genomic features, including a large collection of characterized embryonic enhancers in transgenic embryos (Bonn et al., 2012a; Kvon et al., 2014; Rivera et al., 2019), occupancy of 10 mesoderm/muscle transcription factors profiled in our lab (Cunha et al., 2010; Jakobsen et al., 2007; Junion et al., 2012; Zinzen et al., 2009), DNase I Hypersensitive Sites (DHSs) of FACS purified mesodermal/muscle cells (Reddington et al., 2020), occupancy of 280 transcription factors from the modERN collection (Kudron et al., 2018) and BDGP gene expression data from in-situ hybridization (Tomancak et al., 2002). Sites overlapping a region occupied by Mef2 at 10-12 hr or earlier were classified as Mef2-bound and the non-overlapping sites as unbound. Mef2 motifs were obtained from the collection in (Cannavò et al., 2017) and their presence in sites was scored with function 'AddMotifs' from Signac v1.1.0 (R package). Deciles were calculated with function 'ntile' from dplyr v1.0.2 (R package). Gene expression data over a time course of embryogenesis in *Mef2* -/- embryos was obtained from (Sandmann et al., 2006). Genes were considered

to be associated with Mef2 if a Mef2-bound ATAC-seq peak resided within their gene body or 5kb upstream of their TSS. By this metric, 1,705 genes were associated with at least one Mef2-bound open chromatin region.

3 Optimization of sci-ATAC-seq to profile single-cell chromatin accessibility in *Drosophila* embryos

3.1 sci-ATAC-seq allows to profile single-cell chromatin accessibility at high throughput for thousands of cells

As described in the introduction, several methods have recently emerged to profile chromatin accessibility in single cells (reviewed in (Minnoye et al., 2021)). Among these, single-cell ATAC-seq by combinatorial indexing or sci-ATAC-seq offers high-throughput profiling in the range of tens of thousands of cells per run, and a simple implementation as it does not require high-maintenance microfluidic devices. This method is particularly suited for the profiling of developmental time courses, which require sampling of a large number of cells in order to capture transient developmental transitions. Moreover, thanks to the combinatorial barcoding strategy, sci-ATAC-seq has the flexibility to multiplex several samples in a single run, thus reducing the risk of introducing batch effects among collections from different developmental time points.

A first limitation of sci-ATAC-seq, compared to microfluidic approaches, is that it generally yields a lower cell coverage, meaning that a lower number of sequencing reads are recovered for each single cell. The per-cell coverage is an important parameter of single-cell assays as it essentially represents the sensitivity of the method and determines the resolution of the data that can be obtained. For example, single-cell chromatin accessibility profiling of several human cell lines reached a median coverage of ~ 3,000 reads per cell with sci-ATAC-seq (Cusanovich et al., 2015) and ~ 70,000 reads per cell using a programmable C1 Fluidigm microfluidic device (Buenrostro et al., 2015). A potential reason for this difference is that in microfluidic devices the tagmentation reaction is performed in nanochambers or droplets for each individual cell, which might provide a more efficient reaction environment compared to sci-ATAC-seq where the reaction is performed for thousands of cells simultaneously. More recent applications of sci-ATAC-seq with improved versions of the protocol have been able to achieve a median coverage of ~ 20,000 reads per cell in mouse samples (Cusanovich et al., 2018b; Preissl et al., 2018). A previous study in our lab reached a median coverage of 10,000 reads per cell when applying sci-ATAC-seq to *Drosophila*

melanogaster embryos (Cusanovich et al., 2018a), suggesting that the assay can be further optimized to increase the coverage for *Drosophila*.

A second limitation of sci-ATAC-seq is that it requires a large amount of unloaded Tn5 transposase enzyme to accommodate the 96 unique barcode combinations needed for combinatorial indexing. There are only a few commercial companies providing unloaded Tn5 and the price quickly makes running sci-ATAC-seq cost-prohibitive. Therefore, a homemade version of Tn5 with high capacity for barcode loading and tagmentation needs to be generated. Moreover, the current sci-ATAC-seq protocol utilizes costly Ilumina proprietary reagents for the subsequent steps of library amplification, which make the generation of large sci-ATAC-seq datasets economically challenging. At the beginning of my Ph.D. there was therefore an unmet need to develop a sci-ATAC-seq protocol that utilizes cheaper and more available reagents.

This chapter describes how I optimized sci-ATAC-seq and set up a more cost-effective protocol for the generation of high quality single-cell chromatin accessibility data in *Drosophila melanogaster* embryos.

3.2 Optimization of sci-ATAC-seq for application in *Drosophila melanogaster* embryos

As described above, while sci-ATAC-seq can profile chromatin accessibility at high throughput, a potential limiting factor is the per-cell coverage, i.e. the number of unique reads obtained for each cell. I thus set out to optimize the protocol by identifying reaction conditions that yield increased per-cell coverage in *Drosophila melanogaster* embryos.

I reasoned that there must be tagmentation conditions, in particular the composition of the reaction buffer and the duration of the reaction, that can facilitate the action of the Tn5 transposase and improve its tagmentation activity. A previous study showed that increasing the duration of the tagmentation reaction from 30 minutes to one hour can be beneficial and lead to higher per-cell coverage in mouse cells (Preissl et al., 2018). To verify whether this improvement would be obtained in *Drosophila* samples, I

decided to test reaction times of 30 minutes, one hour and two hours. These were tested in combination with four distinct reaction buffers:

- “Clb”: cold-lysis buffer (clb) is the standard buffer used for sci-ATAC-seq in previous publications (Cusanovich et al., 2018a).
- “Clb fresh”: same buffer as the standard “clb”, but in this version the buffer is prepared fresh just before starting the tagmentation reaction.
- “Omni”: this buffer was developed in the context of Omni-ATAC-seq, a protocol that improves Tn5 tagmentation in frozen tissue samples (Corces et al., 2017). The main feature of “omni” buffer is that it contains additional detergents, in particular digitonin and Tween 20 (also known as Polysorbate 20), to facilitate penetration of the Tn5 enzyme into the nucleus.
- “No dmf”: this buffer has the same composition as “clb” but lacks dimethylformamide (dmf), a commonly used crowding reagent that could negatively affect the Tn5 enzymatic activity.

To conduct the test, I collected *Drosophila melanogaster* embryos at 6-8 hours (mainly stage 11) of development, extracted nuclei and performed sci-ATAC-seq. To avoid batch effects among samples, I exploited the fact that the sample position and the respective tagmentation barcode facilitates multiplexing all the reaction conditions on the same sci-ATAC-seq plate, yielding a total of 12 independent conditions tested in a single run. After sequencing, the number of unique tagmentation reads recovered per cell were quantified for each reaction condition.

I identified a clear effect of time, by which the per-cell coverage steadily increases with the longer duration of the reaction (Figure 5a), a finding that is consistent with previous observations in mouse (Preissl et al., 2018). Extending the reaction duration to two hours increases the per-cell coverage for three of the four buffers tested. Teasing apart the effect of individual reaction buffers is more complex; however, it is clear that the “no dmf” buffer is underperforming in all tested conditions, suggesting that dimethylformamide is in fact helpful to enhance Tn5 enzymatic activity. The “omni” buffer outperforms other buffers at two hours of tagmentation, possibility due to improved nuclear penetration of Tn5 during the extended reaction time. The set of conditions resulting in the highest per-cell coverage (median of 12,481 unique reads

per cell) was achieved using the “omni” buffer with two hours of tagmentation (Figure 5b).

The gain in the per-cell coverage obtained by changing the reaction buffer and introducing a longer tagmentation time could in part be due to an increase in background signal due to non-specific targeting of the Tn5 enzyme to non-accessible regions, thus decreasing the specificity of the signal. To verify this possibility, I computed the number of reads residing within peaks of accessibility previously identified in *Drosophila melanogaster* embryos by bulk DNase-seq (sites here referred to as DHS (DNase Hypersensitive Sites)) (Reddington et al., 2020). For all buffers tested, the fraction of reads residing within open regions decreases with the longer reaction duration (Figure 5c), indicating that more background signal is generated as the reaction time increases. Despite this negative effect, the absolute number of reads residing within open regions is highest at the longest reaction time for most buffers (Figure 5d), indicating that there is still a net gain in specific signal compared with the other reaction conditions.

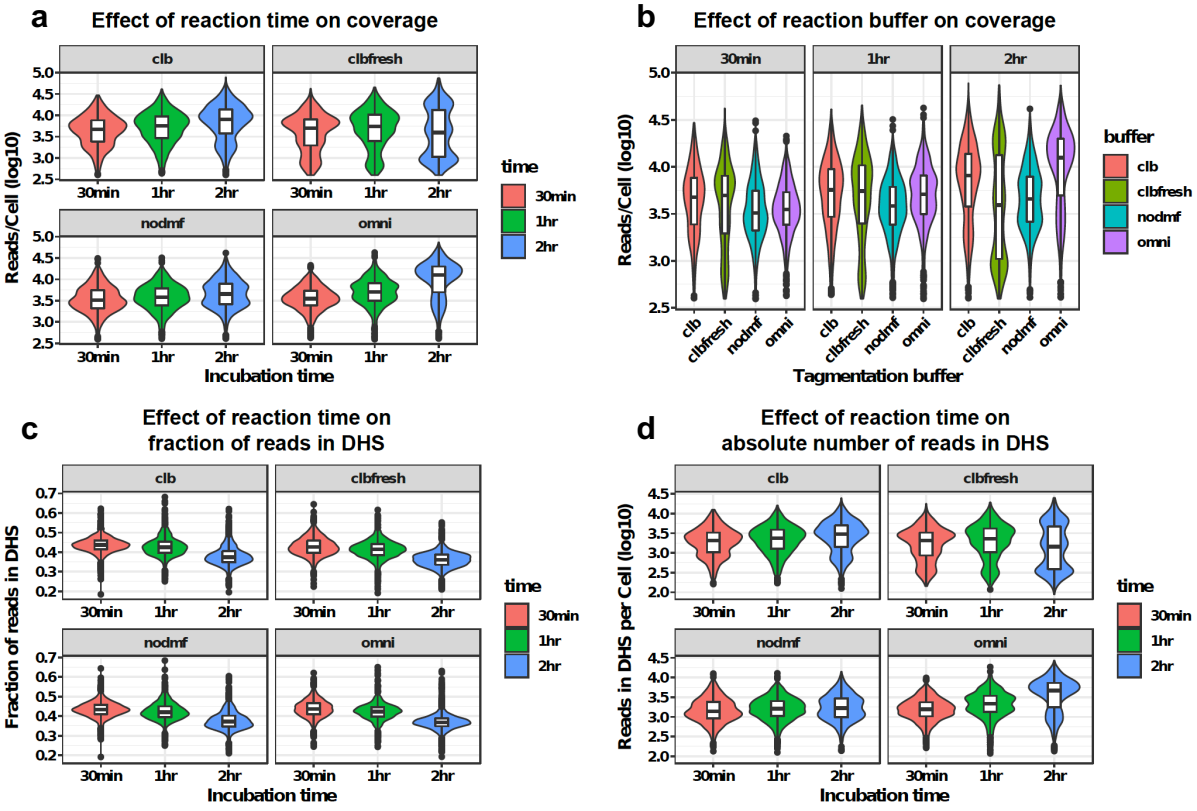

Figure 5. Effect of reaction time and buffer on sci-ATAC-seq per-cell coverage.

(a) Distribution of unique reads per cell for each reaction time split by reaction buffer.
(b) Distribution of unique reads per cell for each reaction buffer split by reaction time.
(c) Distribution of the fraction of unique reads per cell that fall into open (DHS) sites for each reaction time split by reaction buffer.
(d) Distribution of unique reads per cell that fall into open (DHS) sites for each reaction time split by reaction buffer.

Based on these considerations, I decided to generate a full sci-ATAC-seq dataset using the "omni" reaction buffer and allowing the tagmentation reaction to proceed for two hours. These optimizations led to a substantial increase in the assay per-cell coverage. The new dataset yielded a median coverage of 20,108 unique reads per cell, roughly double what was previously reported for *Drosophila melanogaster* (Cusanovich et al., 2018a) and comparable to the median per-cell coverage achieved with the same method in mouse samples (Figure 6a) (Cusanovich et al., 2018b; Lake et al., 2018; Preissl et al., 2018). In comparison with vertebrates, *Drosophila melanogaster* has a considerably smaller genome and therefore, at equal coverage, sci-ATAC-seq captures a larger fraction of the genome (Figure 6b).

My optimized sci-ATAC-seq protocol preferentially captures accessible chromatin regions rather than background, as demonstrated by the higher enrichment obtained by library qPCR targeting a constitutively open region as opposed to a closed one (Figure 6c). These sci-ATAC-seq libraries also show the expected profile of fragment length distribution (Cusanovich et al., 2018a), both in library aggregates and in single cells (Figure 6d, e). Moreover, pseudobulk chromatin accessibility profiles generated by aggregating single cells recapitulate bulk DNase-seq profiles obtained from matched embryonic samples (Figure 6f, bulk DHS data is from (Reddington et al., 2020)). Therefore, the described improvements I made to the sci-ATAC-seq protocol generate high quality libraries with an increased per-cell coverage.

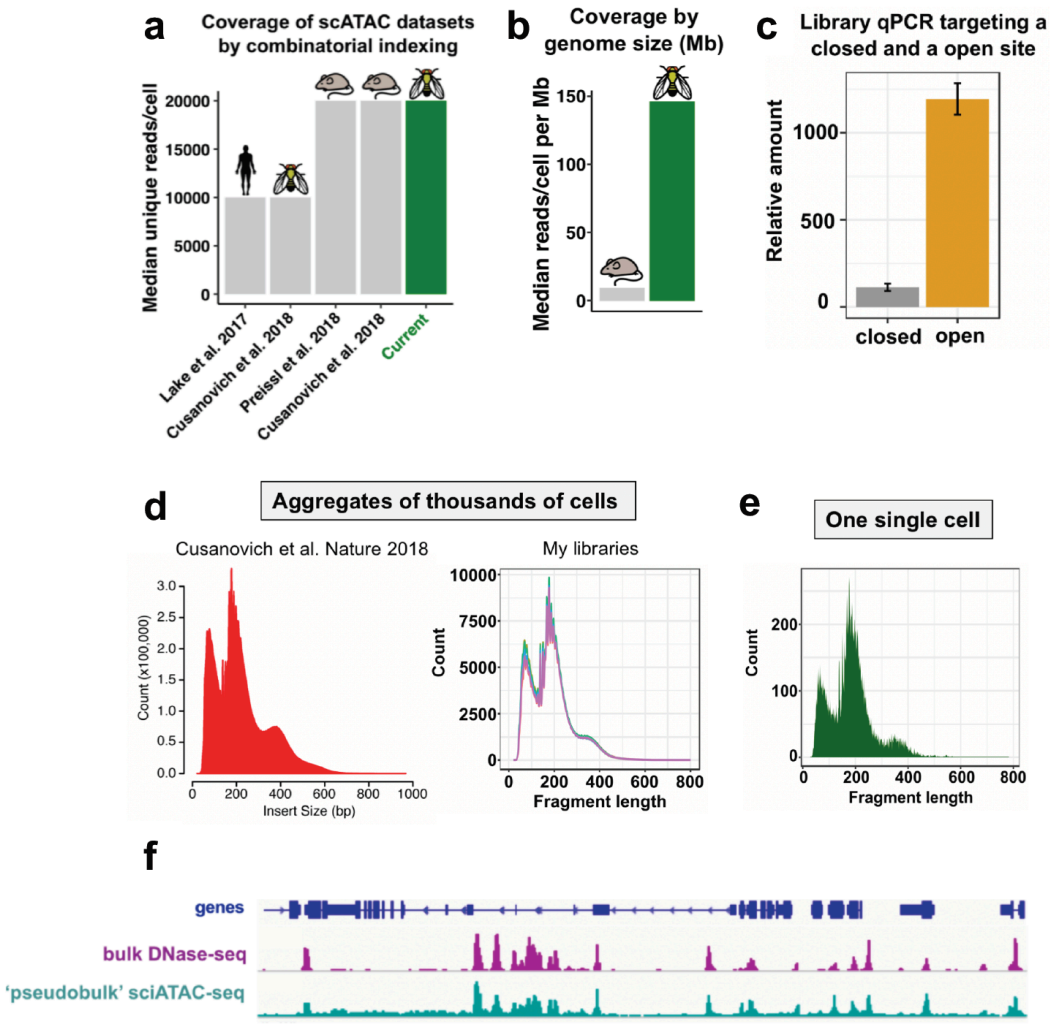

Figure 6. Quality control of sci-ATAC-seq libraries generated with optimized protocol.

(a) Median per-cell coverage (unique reads per cell) obtained in published sci-ATAC-seq datasets for several species and the current experiment.

(b) Median per-cell coverage (unique reads per cell) normalized by megabase of genome for mouse and *Drosophila melanogaster* libraries.

(c) sci-ATAC-seq library qPCR using primers targeting a constitutively closed region and a constitutively open region.

(d) Fragment length distribution of aggregated sci-ATAC-seq libraries. Left: example from a published *Drosophila melanogaster* dataset, right: my dataset. The left panel is adapted from (Cusanovich et al., 2018a) with permission License Number: 5118131129227.

(e) Fragment length distribution for one single cell in my dataset. (f) Genome browser tracks of bulk DNase-seq and aggregated sci-ATAC-seq chromatin accessibility around the Mef2 gene locus. Bulk DNase-seq data is from (Reddington et al., 2020).

3.3 Development of a cost-effective sci-ATAC-seq protocol with homemade reagents

A second limitation of sci-ATAC-seq is that it requires a large amount of unloaded Tn5 transposase enzyme and the protocol is still based on expensive and proprietary commercial reagents. I set out to develop a working protocol using a homemade version of Tn5 transposase coupled with cheaper reagent alternatives.

Tn5 transposase is a bacterial enzyme that catalyzes integration of the Tn5 transposon into the genome (Reznikoff, 2003). The wild-type Tn5 has a generally low rate of transposition. As ATAC-seq requires high insertion frequency in order to profile accessibility genome-wide, mutations have been introduced to generate hyperactive Tn5 versions (Reznikoff, 2008), which are commercially available from companies such as Illumina and Diagenode. However, as sci-ATAC-seq requires 96 uniquely barcoded Tn5 reactions for combinatorial indexing, the cost of using commercial versions for this protocol quickly becomes prohibitive. To circumvent this problem, the EMBL Pepcore facility produced a homemade hyperactive Tn5 using a previously described protocol (Rossi et al., 2018) and made it available for use to our lab. The Tn5 version generated by Illumina has two mutations (E54K L372P) while the version generated by the EMBL Pepcore carries two additional mutations (E54K E110K P242A L372P) that have been shown to reduce the insertion sequence bias (Rossi et al., 2018). To test the performance of the homemade Tn5, I loaded the enzyme with barcodes for combinatorial indexing, following the procedure described in (Preissl et al., 2018), and used it for tagmentation of *Drosophila melanogaster* nuclei, in parallel with commercially available Illumina Tn5. Quantification of the amount of tagmented products by qPCR indicated that the tagmentation efficiency of the homemade Tn5 is the same as the Illumina Tn5.

In addition to the Tn5 transposase, the PCR amplification of sci-ATAC-seq libraries is also cost intensive. The standard protocol utilizes a PCR ready-mix from Illumina called NPM, and the cost of the amount required for one sci-ATAC-seq run is around 3,800 euros. I therefore replaced Illumina NPM with KAPA HiFi ready-mix produced by Roche (catalogue number: 7958935001), which is about ten times cheaper, and used it to amplify sci-ATAC-seq libraries generated with the homemade transposase. Visualization of the tagmented products on high resolution DNA gels indicated that the

fragments distribution is the same for libraries generated with the homemade or the Illumina reagents (Figure 7a). Moreover, sequenced libraries display the expected tagmentation profile (Figure 7b), further confirming that the correct products are generated by the homemade Tn5 and amplified by the KAPA PCR reagent.

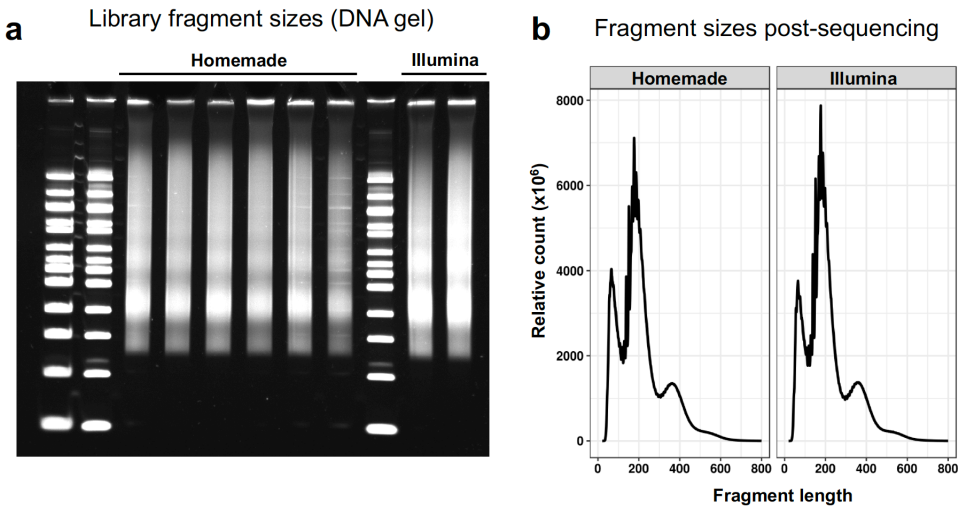

Figure 7. Comparison of tagmentation profiles of sci-ATAC-seq libraries generated with the homemade protocol or the Illumina protocol.

(a) Distribution of tagmented DNA fragments visualized on a DNA gel prior to sequencing.
(b) Distribution of sequenced DNA fragments, displaying the characteristic tagmentation profile of ATAC-seq libraries for both protocols.

The homemade protocol consistently recovers accessible chromatin regions with high specificity. Quantification by qPCR of a constitutively open and closed genomic region demonstrated higher enrichment for the open region, and at a similar enrichment level as for libraries generated with the Illumina protocol (Figure 8a). Aggregated pseudobulk chromatin accessibility profiles for the homemade libraries show clear peaks of accessibility at the same genomic sites as the Illumina libraries (Figure 8b). The homemade libraries have a median coverage of ~10,000 reads per cell, which although lower than I obtained using the Illumina NPM PCR mix, is comparable to the per-cell coverage obtained by Illumina-based sci-ATAC-seq in *Drosophila melanogaster* embryos (i.e. using Illumina Tn5) (Cusanovich et al., 2018a). As demonstrated here, the homemade protocol generates high quality single-cell chromatin accessibility data but is roughly ten times cheaper.

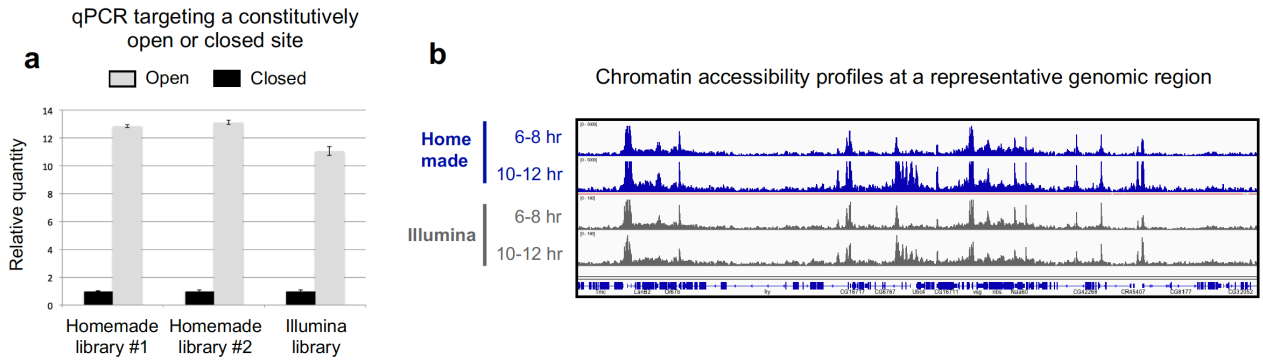

Figure 8. Homemade and Illumina sci-ATAC-seq identify the same regions of chromatin accessibility.

(a) qPCR targeting a constitutively open (grey) and a constitutively closed (black) genomic site for homemade and Illumina sci-ATAC-seq libraries.

(b) Genome browser tracks representing aggregated ('pseudobulk') chromatin accessibility profiles at a representative genomic region for homemade (blue) and Illumina (grey) sci-ATAC-seq libraries at two time points of embryonic development (6-8 and 10-12 hours).

3.4 Discussion and Conclusions

This chapter described how I tackled two outstanding limitations of sci-ATAC-seq, first to increase the assay coverage and second to make the protocol more cost-effective and less dependent on commercial reagents. Thanks to the optimization of the tagmentation reaction conditions, I managed to double the per-cell coverage and I was able to develop a homemade sci-ATAC-seq protocol that is roughly ten times cheaper, allowing for the production of large-scale datasets. Both approaches led to the generation of high quality single-cell chromatin accessibility datasets for *Drosophila melanogaster* embryos. These protocol modifications are not limited to *Drosophila*, but should be compatible with application in a wide range of other species and model organisms.

A first avenue to further improve the per-cell coverage of sci-ATAC-seq would be to continue exploring changes to the buffer conditions for the tagmentation reaction. At present, I have tested four distinct buffers that are based on already existing recipes. A more systematic approach by performing a large screen for a whole library of

compounds, detergents and other buffer additives to assess their impact on Tn5 activity could yield more efficient tagmentation. As sci-ATAC-seq is performed in 96-well plates, it is amenable to testing multiple conditions simultaneously and also suitable for further upscaling with the use of robot handlers. The Tn5 activity in the different reaction environments could be monitored by quantifying the tagmentation products by qPCR. As the rate of transposition is determined by the activity of the Tn5 enzyme, a second potential avenue for improvement would be to generate new versions of Tn5 transposase with additional mutations and test if they confer higher tagmentation activity. This approach would however be time consuming and less practical as it requires the production and isolation of the Tn5 variants in bacterial cultures.

Finally, while the homemade sci-ATAC-seq protocol generates high quality data, changing the PCR components led to a decrease in the per-cell coverage, when compared to the use of Illumina PCR reagents. More reaction mixes should be tested to identify the ones with higher polymerase activity that lead to the same or higher coverage than the standard protocol.

3.5 Contributions

I performed all the experiments and computational analysis presented in this chapter, unless otherwise stated.

Staff at the EMBL Protein Expression and Purification Facility produced the homemade Tn5 enzyme.

Staff at the EMBL Genecore Facility carried out all sequencing. I prepared all the libraries myself, and submitted them to the facility for sequencing.

Dr. Eileen Furlong (*Group Leader, Furlong lab*) supervised the project, made intellectual contributions and provided funding.

4 Using sci-ATAC-seq to phenotype mutants at both a cellular and molecular level

Understanding the progression and regulation of cell lineages is a longstanding goal of developmental biology. As discussed in the introduction, the implementation of new single-cell technologies such as scRNA-seq and scATAC-seq enables to probe developmental processes with unprecedented resolution. Single-cell RNA-seq, applied to multiple developmental stages, provides a powerful approach to uncover cellular diversity, identify new cellular states, and chart lineage trajectories during embryonic development, as demonstrated for a range of different organisms (Briggs et al., 2018; Davie et al., 2018; Farrell et al., 2018; Özel et al., 2021; Pijuan-Sala et al., 2019; Plass et al., 2018; Tyser et al., 2021). Decoding how regulatory networks drive the development of embryonic lineages requires an understanding of both the TFs involved and the enhancers they bind to. While scRNA-seq can identify which TFs are expressed along different developmental lineages, it provides no information on the regulatory elements, e.g. which enhancers they act through. Single-cell regulatory genomics methods (Minnoye et al., 2021), such as single-cell ATAC-seq, provide a direct approach to measure enhancer usage across different tissues and stages of embryogenesis, as recently demonstrated in *Drosophila* (Cusanovich et al., 2018a), mouse (Lake et al., 2018; Pijuan-Sala et al., 2020), and human (Domcke et al., 2020) embryogenesis.

While these studies have been a great step forward, they are limited to profiling relatively narrow temporal windows in most cases, and therefore do not reveal enhancer usage over an entire, continuous lineage's development. In addition, most studies only profiled wild-type samples. The components of regulatory networks need to be perturbed to gain a functional understating of how they operate. Traditional genetic studies addressed this by mutating key TFs in a regulatory network and then assessing the phenotype at either the cellular level, describing high-level tissue abnormalities by immuno-staining, and then separately at the molecular level by performing RNA-seq and/or ChIP-seq, with limited integration between the two. The development of single-cell genomics methods provides a new opportunity to change this. Profiling a high-resolution time-course that densely covers the development of a

tissue in wild-type embryos, combined with loss-of-function TF mutants should be able to provide a detailed view of the functional role of developmental factors at both a cellular and molecular level.

The specification of the mesoderm into different muscle primordia in *Drosophila melanogaster* offers a well-studied model system to explore this further (Azpiazu and Frasch, 1993; Azpiazu et al., 1996; Riechmann et al., 1997; Ruiz-Gómez, 1998; Sandmann et al., 2007). This germ-layer gives rise to all major muscle types from flies to humans, and the key TFs regulating the subdivision of the mesoderm into different muscle lineages are known and highly conserved (Ciglar and Furlong, 2009). Seminal genetic screens in *Drosophila* uncovered the functional requirement of many of these factors in muscle development, describing high-level phenotypes such as missing or abnormal muscles (Azpiazu and Frasch, 1993; Bour et al., 1995; Lilly et al., 1995; Zaffran et al., 2001). The Furlong lab has also generated extensive molecular data (ChIP, RNA, DNAse) in wild-type embryos describing TF binding and enhancer usage during mesoderm and muscle development (Junion et al., 2012; Reddington et al., 2020; Zinzen et al., 2009).

In this chapter, I describe a general framework to better integrate these two, i.e. the high-level tissue phenotypes with the molecular input of key TFs. I first generated a dense time-course of single-cell regulatory changes during mesoderm development in wild-type *Drosophila* embryos, using the new improvements I made to the sci-ATAC-seq protocol described in the previous results chapter. I performed sci-ATAC-seq on FAC sorted mesodermal nuclei across eight overlapping embryonic time-points, capturing a continuum of regulatory transitions as cells move from multipotency to different developmental lineages. I then used this high-resolution dataset to identify new regulatory elements and transcription factors, and to reconstruct regulatory trajectories of each muscle lineage. I then exploit these trajectories to examine the phenotypes of four essential TFs using loss-of-function mutants, providing a high-resolution view of their functional impact at a cellular (where are the mutant cells located on the wildtype trajectory) and at a molecular level (which enhancers and genes are affected). I demonstrate that this approach can pinpoint known tissue defects *de novo*, and uncover more subtle phenotypes that may have been missed by

immuno-stains, in addition to giving new insights into the regulatory nature of the transcription factor.

In this chapter, I performed both the experimental work to generate the sci-ATAC-seq data and the new isogenic CRISPR mutants, as well as all of the sci-ATAC-seq data analyses. The only exception being the nuclear genotyping – this analysis was done by Tobias Heinen and Mattia Forneris.

4.1 Capturing single-cell chromatin accessibility during a comprehensive time-course of mesoderm development

As described above, while single-cell ATAC-seq has been applied to profile embryonic development, the studies so far have been quite limited in their temporal resolution, thus preventing a continuous reconstruction of developmental trajectories. *Drosophila* has an advantage here as it is relatively easy to obtain tightly staged embryos (within hours compared to half days or days within some organisms). Taking advantage of this, I profiled chromatin accessibility by sci-ATAC-seq in eight overlapping, rather than adjacent, 2-hour tightly staged embryo collections to resolve continuous single-cell trajectories during mesoderm development (Figure 9). This time course initiates shortly after gastrulation when mesodermal cells are still multipotent (3-5hr), and continues through the stages of cell fate specification (~6-8hr) to terminal tissue differentiation (10-12hr), ensuring that all major developmental transitions are captured and can therefore be followed. For each two-hour staged collection, I formaldehyde-fixed *Drosophila melanogaster* embryos, extracted nuclei and FAC sorted them with a mesodermal/muscle marker (Mef2) to capture the developing muscle lineages using a previous protocol developed in the Furlong lab (Reddington et al., 2020) (Figure 9).

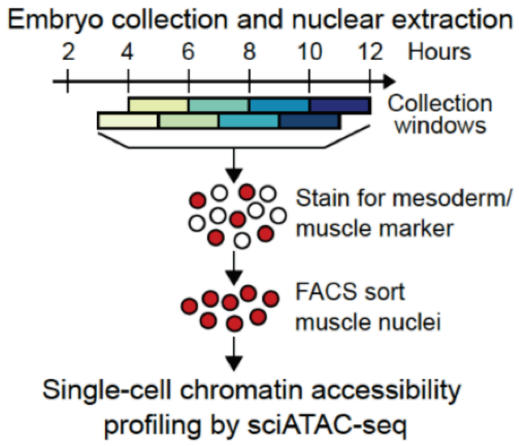

Figure 9. Illustration of the experimental design.

Staged *Drosophila melanogaster* embryos were collected from eight overlapping 2-hour windows and mesoderm/muscle nuclei were FAC sorted based on the Mef2 marker to >95% purity prior to sci-ATAC-seq.

I performed two sci-ATAC-seq 'replicates' per time-point, consisting of independent embryo collections and mesodermal nuclei sorting, and independent days where the sci-ATAC-seq protocol was run. This resulted in a combined dataset of 24,032 single cells that passed standard sc-ATAC quality filters, based on unique read-depth per cell and nucleosomal patterning (for details please refer to Materials and Methods section 4.5.2: 'Barcode-cell assignment and quality control'). Through the protocol optimizations discussed in the first results chapter (Chapter 3), I roughly doubled the number of reads per cell compared to a previous study produced by our lab (Cusanovich et al., 2018a), obtaining a median of 21,649 unique reads per cell. All indicators suggest that the sci-ATAC-seq dataset that I generated (the wild-type mesoderm trajectory) is of very high quality: pseudo-bulk chromatin accessibility profiles, created by aggregating the single cell data, are (1) highly correlated among replicate sci-ATAC-seq batches (Pearson correlation coefficient r = 0.98), they (2) recapitulate bulk DNase-seq profiles at muscle enhancers with high tissue specificity (Figure 10a) and globally they are (3) highly correlated with bulk DNase-seq profiles from time-matched FAC sorted muscle populations (Figure 10b).

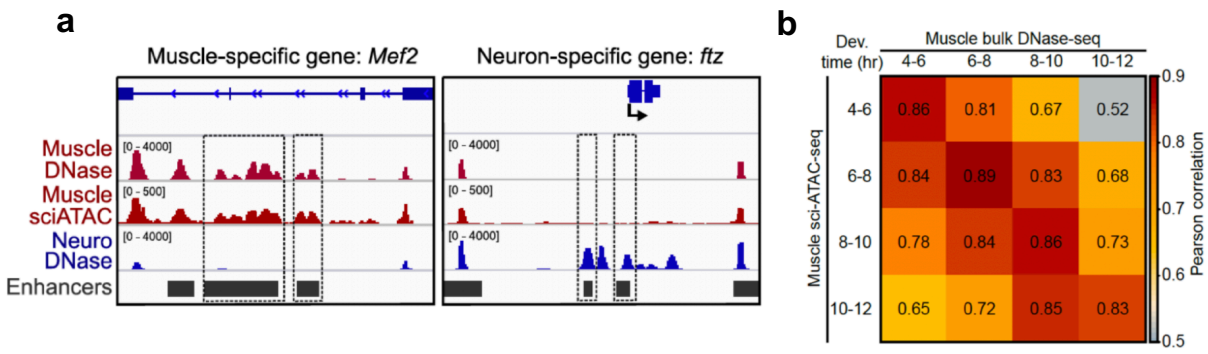

Figure 10. Quality assessment of the single-cell ATAC-seq profiles of mesoderm development.

(a) Bulk DNase-seq and aggregated sci-ATAC-seq chromatin accessibility profiles at a muscle-specific (left - *Mef2*) and neuron-specific (right - *ftz*) gene. Characterized developmental enhancers are indicated in grey.

(b) Correlation between bulk DNase-seq accessibility profiles from FAC sorted mesoderm/muscle cells and aggregated pseudobulk sci-ATAC-seq profiles for time matched embryo samples.

The DNase-seq data used in panel (a) and (b) was retrieved from (Reddington et al., 2020).

4.2 The single-cell chromatin accessibility landscape of embryonic muscle development

Analysis of single-cell chromatin accessibility data is challenging, due to the noise generated by the inherent data sparsity (low coverage per cell) and the high dimensionality of the datasets. To tackle this challenge, dimensionality reduction methods that can transform noisy high-dimensional single-cell data into more interpretable low-dimensional representations have been developed (Sun et al., 2019). Among the available tools for unsupervised clustering of single-cell ATAC-seq data, I decided to use cisTopic (Bravo González-Blas et al., 2019), because it outperforms other clustering methods when applied to continuous populations (Chen et al., 2019).

To resolve the chromatin accessibility landscape, I applied cisTopic to cluster single cells based on their chromatin accessibility profiles and yield a representation of the dataset that I then visualized in a UMAP dimensionality reduced space (for details

please refer to Materials and Methods section 4.6.2: ‘LDA, clustering and UMAP visualization’). This unsupervised clustering of all cells from all time-points revealed a tree-like structure that reflects the temporal order of the embryonic collections, revealing the dynamic changes in chromatin accessibility as the development of the mesoderm unfolds (Figure 11). Cells sorted from the early time-points are relatively uniform in their regulatory landscape and form a trunk that represents the nascent unspecified mesoderm (Figure 11, light yellow). The chromatin accessibility landscape diversifies at around 6-8 hours of embryogenesis, resulting in branches along different trajectories (Figure 11). This diversification at 6-8 hours (stages 10-11) matches the time window uncovered by genetic studies in the 1990s for the subdivision of the mesoderm into different muscle primordia (Azpiazu and Frasch, 1993; Azpiazu et al., 1996; Riechmann et al., 1997). The clustering algorithm had no prior information about this, and yet this naturally falls out from the data.

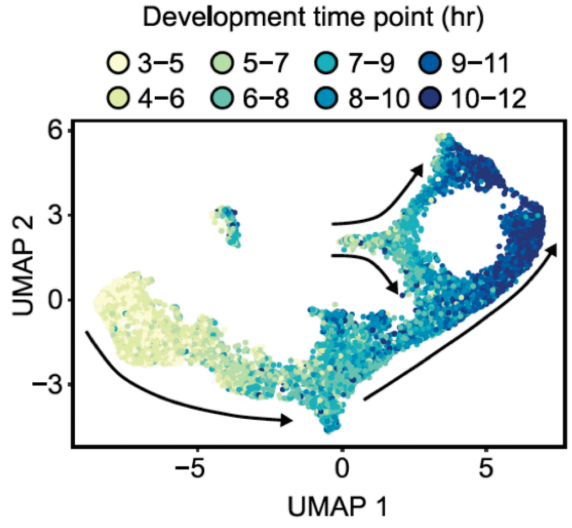

Figure 11. The regulatory landscape of embryonic muscle development.

UMAP visualization of all the cells profiled over the whole time course. Each dot represents a cell, colored by the time point of the embryo collections. Cells are clustered based on the similarity in their chromatin accessibility.

4.3 Cell-type identification reveals that the muscle lineages have distinct chromatin accessibility profiles

A major challenge in single cell genomes and large-scale atlas projects is to identify the cell types in different clusters. To do that within my mesoderm/muscle atlas, I used Seurat (Stuart et al., 2019) defined clusters, which partitioned the dataset into fifteen major cell clusters (Figure 12a) and looked for over-representation of tissue-terms within each cluster (for details please refer to Materials and Methods section 4.7: 'Cell-type annotation') using two resources: (1) characterized embryonic enhancers (Bonn et al., 2012a; Kvon et al., 2014; Rivera et al., 2019) and (2) gene expression data throughout embryogenesis (Tomancak et al., 2002). These independent approaches resulted in highly concordant annotations that allowed me to identify the cell-types in each cluster (Figure 12b, Figure 13a).

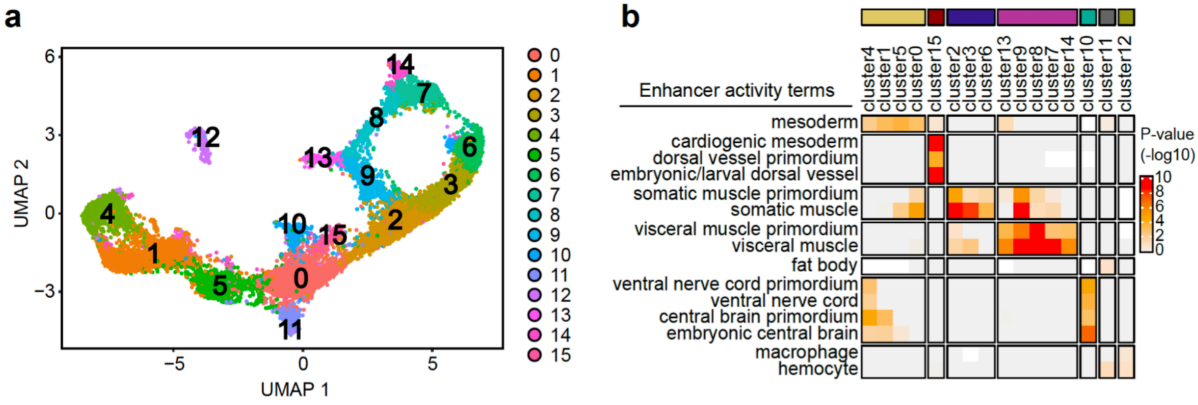

Figure 12. Annotation of cell clusters in the mesoderm/muscle time course.

(a) UMAP visualization of the full mesoderm/muscle time course (same as Figure 11), cells are colored by their respective Seurat defined cell cluster.

(b) Enrichment for enhancers with embryonic tissue activity terms in each cluster. Fisher's exact test P-value (-log10) displayed as a heatmap. Cell clusters are grouped by the inferred cell-type annotations (top bars). The annotation colors correspond to those displayed in Figure 13a.

This process resolved the early mesoderm population and the three major myogenic lineages: the somatic, visceral and cardiac muscles (Figure 13a), indicating that they have distinct chromatin accessibility landscapes and require differential usage of

enhancers and promoters. As specification begins half way through the time course, unspecified mesodermal cells constitute the largest population of cells, while the cardiogenic mesoderm and resulting heart muscle is the least abundant (Figure 13a). The fact that I can detect cardiomyocytes, a rare cell population that represents roughly 0.2-0.3% of the embryo at the end of embryogenesis (Reim and Frasch, 2010), suggests that I have comprehensively sampled the diversity of myogenic cell types at these stages. The continuum of muscle development is reflected in the distribution of different cell populations, which shifts over developmental time with the early uncommitted muscle progenitors being gradually replaced by the terminal muscle types (Figure 13b). I also identified three small non-myogenic populations. Fat body and hemocytes are both derived from the mesoderm and accordingly are present at the specification stages but absent at later time points, as *Mef2* expression becomes more restricted to myogenic populations (Figure 13b). The third non-myogenic population are neural cells, which most likely come from a subpopulation of *Mef2* expressing cells within the mushroom body of the brain (Crittenden et al., 2018).

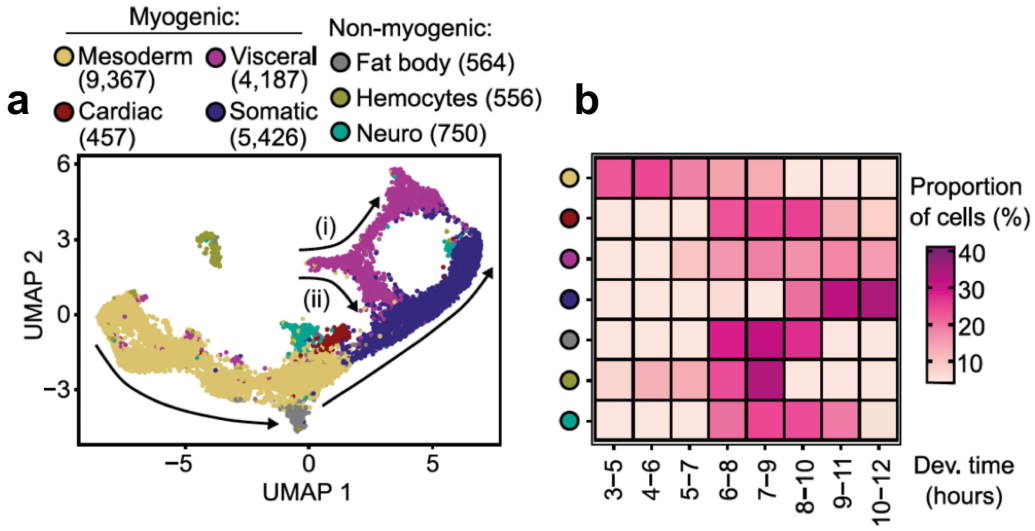

Figure 13. Developing muscle lineages have distinct chromatin accessibility profiles.

(a) UMAP visualization of the full mesoderm/muscle time course (same as Figure 11), cells are colored by their cell-type annotation. The number of cells of each annotation is indicated in parenthesis.

(b) Percentage of cells in each developmental time point (hours) by cell population.

To experimentally validate the cell-type assignment, I FAC sorted cells from one specific muscle population, the visceral muscle using the lineage-specific marker Biniou, and obtained high-quality sci-ATAC-seq profiles for 1,295 sorted visceral muscle cells. These cells were actually clustered with the Mef2 sorted cells shown in the UMAPs in Figures 11-13. Highlighting where these visceral muscle cells are on the UMAP shows that they cluster together with the cells assigned as visceral muscle (Figure 14a (orange) and Figure 13a (purple)). To further confirm the cell cluster annotations, I looked for known marker genes for lineage specification and differentiation. All show high accessibility in the expected muscle populations at the appropriate stages further confirming the accuracy of the annotations (Figure 14b); including *tinman* (*tin*), *pannier* (*pnr*) and *Doc3* (*Doc3*) in the cardiac muscle, *bagpipe* (*bap*), *biniou* (*bin*) and *Shaker cognate 1* (*Sha1*) in the visceral muscle, *Muscle-specific protein 300 kDa* (*Msp300*) and the contractile proteins *Tropomyosin 1* (*Tm1*) and *Myosin heavy chain* (*Mhc*) in differentiated somatic muscle.

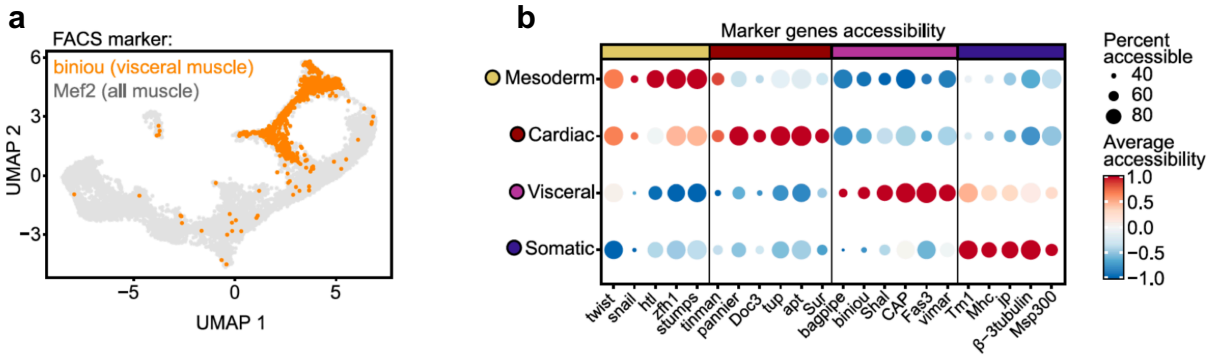

Figure 14. Validation of inferred cell-type identities.

(a) UMAP visualization of the full mesoderm/muscle time course (same as Figure 11). Grey cells were FAC sorted using a general mesoderm/muscle marker (Mef2), orange cells were isolated using a visceral muscle specific marker (Biniou).

(b) Dotplot of marker genes accessibility in each muscle cell-type. The color scale indicates the gene average accessibility (Z-score), while the dot size indicates the percentage of cells in which the gene is accessible.

4.4 Chromatin accessibility changes along muscle development reflect dynamic transcription factor activity

To resolve TFs that might be responsible for the regulatory changes along mesoderm development, I first examined dynamic changes in accessibility at regions bound by muscle specific factors by calculating TF deviation scores (Schep et al., 2017). These scores are indicative of the level of accessibility at the regions bound by a given TF, which is used to infer the TF's dynamic activity. The TF activities inferred with this approach match the expected patterns for each given TF (Figure 15a). For example, *tinman* is broadly expressed throughout the trunk mesoderm at early stages before being restricted to the dorsal mesoderm and cardiac muscle at later stages of embryogenesis (Azpiazu and Frasch, 1993; Yin and Frasch, 1998). Concordantly, Tinman bound sites from bulk ChIP data at 2-4 hours and 4-6 hours show high accessibility in single cells of the early mesoderm, while Tinman binding at 6-8 hr is restricted to single cells assigned as cardiac muscle (Figure 15a). Similarly, Biniou and Bagpipe, two factors required for visceral mesoderm specification (Azpiazu and Frasch, 1993; Zaffran et al., 2001), show specific activity in the visceral muscle single-cell cluster (Figure 15b). Cells with open chromatin sites that are bound by the pan-muscle factor Mef2 at either 2-4, 6-8 or 10-12 hours of development display concordant changes in the accessibility of these sites over time, with the sites being more accessible at early, mid and late points in my single-cell time course, respectively (Figure 15b, upper). Similar temporal specific activity is seen for cells with regions overlapping Twist binding at 4-6 hr and Biniou and Lame-duck binding and 6-8 hr (Figure 15b, lower). Taken together, this indicates that my single-cell chromatin accessibility atlas accurately recapitulates the underlying temporal and spatial patterns of TF activity during mesoderm development.

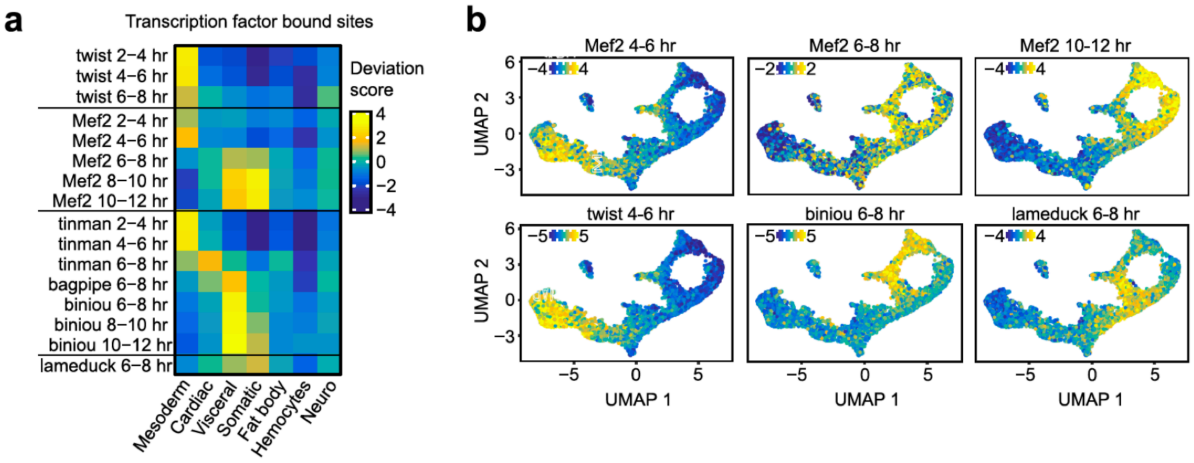

Figure 15. Dynamic activity of mesodermal TFs is reflected in chromatin accessibility changes along muscle development.

(a) Heatmap of accessibility deviation scores (calculated with chromVAR (Schep et al., 2017)) for transcription factor bound regions (using bulk ChIP data) per cell population.

(b) UMAP plots with cells colored by the accessibility deviation score for the indicated transcription factor using bulk ChIP data from specific time-points (as in panel a).

The bulk ChIP data used in panel (a) and (b) was retrieved from 16 datasets of mesodermal TFs produced by our lab (Cunha et al., 2010; Jakobsen et al., 2007; Junion et al., 2012; Zinzen et al., 2009).

To identify additional potential regulators of different muscle sub-types, I computed deviation scores for a large collection of 280 TFs active during embryogenesis (Kudron et al., 2018). This bulk ChIP-seq data was generated from whole-embryos collected over very large developmental time windows spanning either half or all of embryogenesis, and thereby averages signal over different cell-types and time-points. Nevertheless, the high resolution of my single-cell data could resolve both the cell-type and rough time window of occupancy of several factors (Figure 16). For example, resolving whole-embryo ChIP data for Tinman and Tailup to the cardiac muscle (Azpiazu and Frasch, 1993; Tao et al., 2007; Zmojdzian and Jagla, 2013), Nautilus and Pdp1 to the somatic and visceral muscle (Abmayr and Keller, 1997; Lin et al., 1997), and Org-1 and FoxL1 to visceral muscle (Hanlon and Andrew, 2016; Schaub and Frasch, 2013), consistent with the role of these TFs in the corresponding cell types. Nau is a good example of refining the temporal-window - the ChIP data was performed on whole embryos from a very broad time-window spanning almost all of

embryogenesis (4-24hr). However, integration with my single cell time-course could resolve the time-window to 10-12hrs, and the tissue to the somatic and visceral muscle (Figure 16).

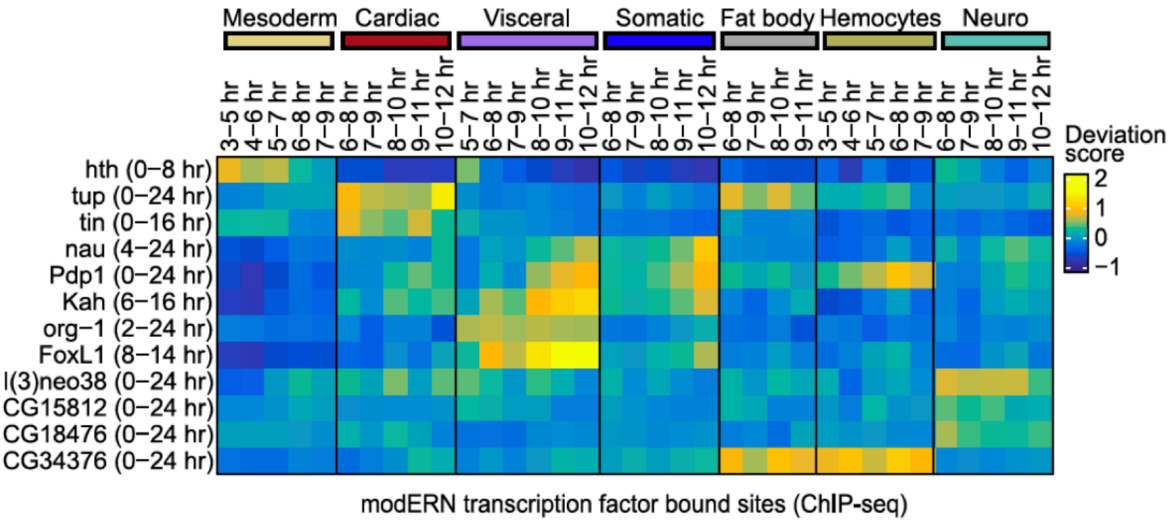

Figure 16. Refining TF activity from occupancy data profiled at low temporal and spatial resolution.

Heatmap of accessibility deviation scores (calculated with chromVAR (Schep et al., 2017)) by cell population and developmental time (hours) for sites occupied by selected transcription factors. Time points containing less than 5% of cells of a given population were excluded. The bulk ChIP-seq data of each TF was retrieved from (Kudron et al., 2018).

4.5 Identification of new putative enhancers and regulators in each muscle lineage

I next used this high-resolution regulatory atlas to identify new genes and putative enhancers that are differentially accessible in a specific tissue or muscle sub-population (Figure 17a, (Data Table 1)). Of the 5,180 differential, i.e. tissue-specific, peaks (representing 36% of all tested peaks), 78% (4027/5180) are distal from an annotated promoter and are likely enhancers. In support of this, 20% (790/4027) overlap previously characterized embryonic enhancers *in vivo*. 19% (752/4027) of this set of distal differential elements are novel, and were not discovered in previous whole embryo shot-gun sci-ATAC-seq (Cusanovich et al., 2018a) or tissue-specific DNase-

seq (Reddington et al., 2020) studies (Figure 17b), which is quite remarkable given that this system has been profiled by many bulk genomic studies.

In addition to distal elements, 864 genes are differentially accessible across their gene body between cell types (Figure 17a, (Data Table 1)), which can serve as a proxy for changes in gene expression (Granja et al., 2021). Many of these are components of signaling pathways or TFs, including many known regulators of mesoderm/muscle development, thus validating the approach. For example, the transcription factors *pnr*, *Doc3*, *tup* and *apt* in the cardiac muscle, *bap*, *bin*, *H2.0*, and *hand* in the visceral muscle and *Pdp1* and *cf2* in the somatic muscle (Data Table 1). In addition to these well characterized TFs, the data also uncovers many new regulators with differential accessibility in specific tissues (Data Table 1).

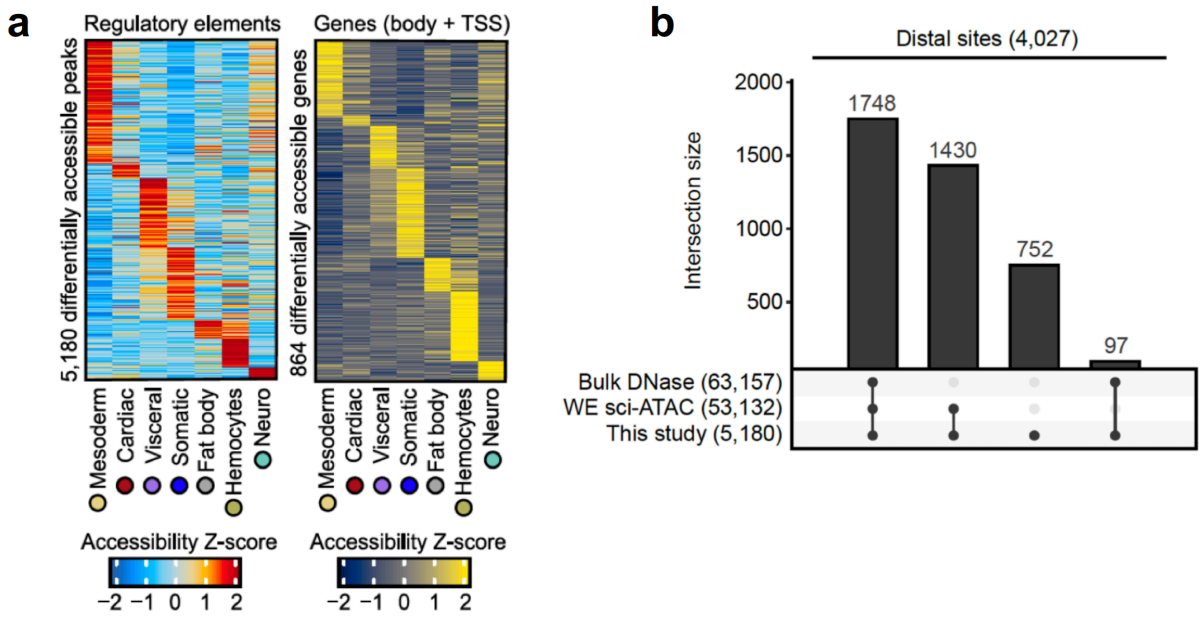

Figure 17. Identification of differentially accessible regulatory elements and genes in each muscle lineage.

(a) Heatmaps of average accessibility (Z-score) at differentially accessible ATAC-seq peaks (left, regulatory regions) and genes (right) in each cell population.

(b) Upset plot showing the intersection between differentially accessible (DA) distal ATAC-seq peaks discovered in this study (4,027 distal peaks out of 5,180 DA peaks from panel (a)) compared to 63,157 peaks from tissue-specific bulk DNase-seq (Reddington et al., 2020) and 53,132 peaks from whole-embryo (WE) shot-gun sci-ATAC-seq (Cusanovich et al., 2018a).

DA sites are classified as distal if they reside > 500 bp away from a Transcriptional Start Site (TSS). An intersection is counted if at least 25% of a peak in this study is overlapped by a peak in another dataset.

4.6 Dynamic changes in regulatory elements are sufficient to reconstruct diverse lineage trajectories

I next exploited the continuous temporal resolution of my single-cell time-course to reconstruct regulatory trajectories for the development of diverse muscle types, beginning from unspecified mesodermal cells in early embryogenesis (Figure 18a, yellow dot) to different end points, reflecting the development of different muscle lineages. Ordering cells along pseudotime using ArchR (Granja et al., 2021) uncovered widespread and very dynamic temporal changes in accessibility for both distal regulatory elements and genes along each lineage's trajectory (Figure 18b). The loci of many cellular identity genes including TFs change in accessibility (in both directions) as the development of each lineage proceeds. This includes *lmd*, *kah*, *NK7.1*, *Pdp1*, *tx*, *nau (dMyoD)* in the somatic trajectory, in addition to more downstream effector genes required for differentiated muscle function (e.g. the contractile proteins *Mhc*, *Mlc*, *Tropomyosin*). The heart cardiomyocytes are specified by a highly conserved set of TFs from flies to humans (Davidson and Douglas, 2006), including members of the NKx2.5 (*tinman* (*tin*) in *Drosophila*), GATA (*pannier* (*pnr*)), T-box (*Dorsal corss-3* (*Doc3*)) and islet 1 (*tailup* (*tup*)) TFs. The dynamic usage of all these factors and many more are observed along the cardiac lineage (Figure 18b).

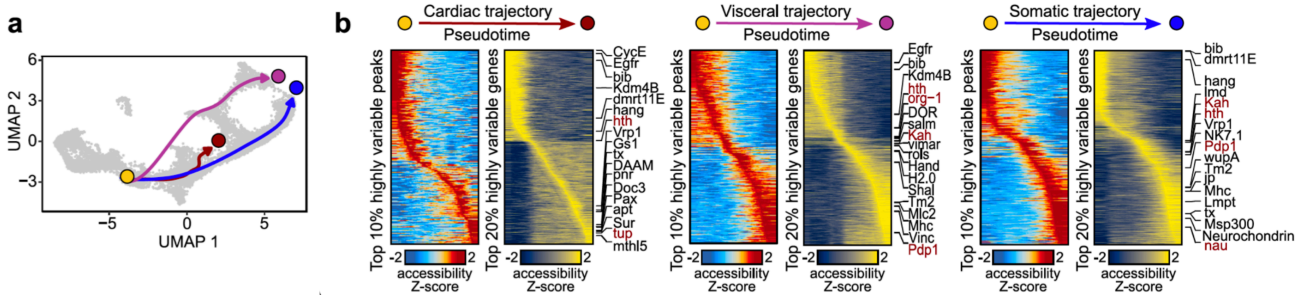

Figure 18. Reconstruction of lineage trajectories for each muscle type.

(a) Inferred cardiac (red), visceral (pink) and somatic (blue) developmental trajectories on UMAP starting from a common point in the unspecified mesoderm (yellow dot).
(b) Heatmaps of ATAC-seq peaks and genes accessibility through pseudotime for each trajectory. The top 10% and 20% most highly variable peaks and genes across pseudotime are visualized. Transcription factors identified in Figure 16 are colored in red.

4.7 Uncovering the developmental progression of visceral muscle sub-populations

Both the somatic muscle and visceral muscle are formed from two populations of cells (each with sub-populations) – founder cells (FCs), which give the muscle its identity, and fusion competent myoblasts (FCMs), which fuse to FCs during muscle differentiation to form a multi-nucleated syncytium (Deng et al., 2017; Lee and Chen, 2019). I observed two trajectories for the visceral muscle lineage (Figure 13, purple cells). Starting from common precursors, one lineage branches towards the somatic body wall muscle (Figure 13 (ii)), while the other has a distinctive lineage that remains separate from the rest of the sub-muscle populations (Figure 13 (i)). To explore the visceral muscle lineages further, I clustered all visceral muscle cells separately, which revealed a more complex multi-branched structure (Figure 19a). The visceral muscle represents a collection of muscles with different developmental origins (Figure 19a, embryo scheme). The FCs and FCMs of the circular trunk visceral muscle (CVM) are specified at stages 10/11 (6-8 hr), after which they migrate laterally and undergo myoblast fusion (stage 12) to form a continuous muscle that encloses the gut (Lee et al., 2006). The longitudinal VM (LVM) is formed from FCs that originate in caudal mesoderm towards the end of the germband, which migrate on top of the circular VM, using it as a scaffold (Zaffran et al., 2001).

My single-cell trajectory captures these diverse origins and the temporal delay of the LVM development. The main branch (Figure 19a, cluster 0) consists of both FCs and FCMs, as seen by their enrichment in markers for visceral muscle FC specification (including *Alk*, *numb*, *bap*, *bin*) and FCM specific genes (e.g. *lmd*, *sns*) (Figure 19b). A

subset of cells progress to a more differentiated state (Figure 19a, moving from cluster 0 → 1→ 2), expressing specific myoblast fusion genes in the intermediate state before differentiating to contractile muscle, indicated by the expression of sarcomere protein genes such as *MHC*, *Tmod*, *Tm1* and *Tm2* (Figure 19b). Given the timing of these transitions and the presence of *bap* (a TF not expressed in LVM) this very likely reflects circular VM development. The LVM and hindgut VM (HVM) branches have different developmental origins, as expected, and are enriched in longitudinal (cluster 5) and hindgut (cluster 4) visceral muscle markers (Bae et al., 2017; San Martin and Bate, 2001) (Figure 19b). The precursor population has a second branch (Figure 19a, cluster 0 → 3), expressing a broader set of myoblast fusion genes (Figure 19b), maybe representing a more mature or more diverse population of FCMs. It has been proposed that some FCMs remain unfused until the LVM founder cells migrate over the circular VM and then fuse (during late stage 12, stage 13 (~10-12hr)) (Klapper et al., 2002). My results support this, and suggest that a proportion of these cells may also fuse to the overlying somatic muscle FCs, given their expression of *Vrp1/wip* and *wasp*, two genes specific to myoblast fusion of the somatic muscle (Rudolf et al., 2014).

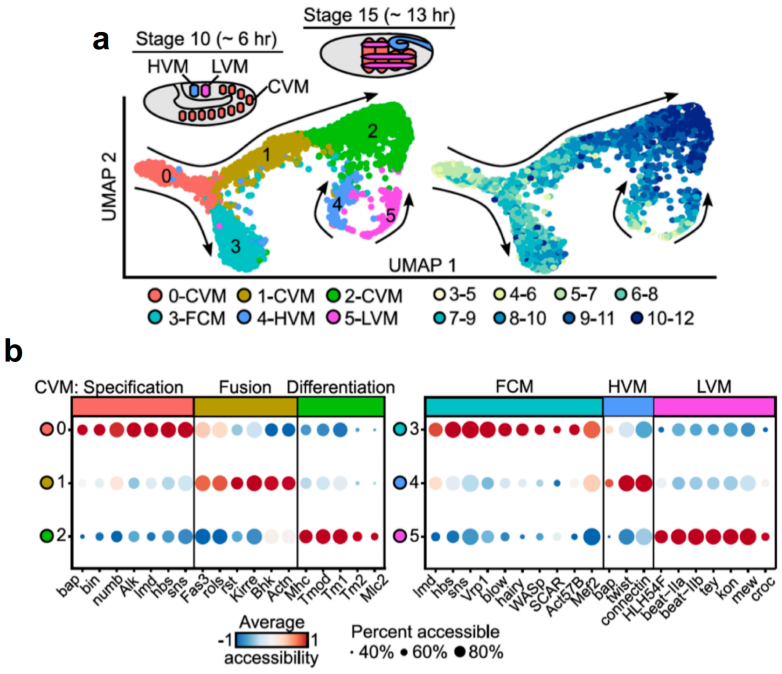

Figure 19. Re-clustering the visceral muscle reveals heterogenous sub-populations.

(a) Top: schematic representing the development of the major embryonic visceral muscle (VM) sub-types. CVM: Circular VM, FCM: Fusion Competent Myoblasts. LVM: Longitudinal VM, HVM: Hindgut VM. Bottom: UMAP visualization of the re-clustered visceral muscle population

(purple cells from Figure 13), cells colored by cluster (left) and developmental time point (hours) (right).

(b) Marker genes accessibility for CVM developmental progression (left) and other VM sub-types (right). The color scale indicates the gene average accessibility (Z-score), while the dot size indicates the percentage of cells in which the gene is accessible.

4.8 Fast and streamlined generation of mutant data by single-nucleus *de-novo* genotyping

The genotyping was done in collaboration with Tobias Heinen and Mattia Forneris. Their individual contributions are indicated below.

To determine the functional impact of loss-of-function TF mutants at the level of regulatory programs and developmental trajectories, I applied sci-ATAC-seq to loss-of-function mutant embryos of four mesodermal TFs, and assessed the outcome by integrating them with the wild-type developmental trajectory (Figure 13). *Mef2* (Myocyte Enhancer Factor 2) is essential for myoblast fusion and terminal differentiation of all muscle types. *Tinman* (Nkx2-5) is essential for the subdivision of the dorsal mesoderm into cardiac and visceral muscle cell fates, through the activation of *bagpipe* (NKx3-2), which initiates *biniou* (FoxF2) expression and the visceral muscle lineage. Although the occupancy of each TF has been examined in bulk (Jakobsen et al., 2007; Junion et al., 2012; Liu et al., 2009; Sandmann et al., 2006, 2007; Zinzen et al., 2009), their contribution to enhancer accessibility, and to an individual cell's state, remains unknown.

Recessive lethal mutants in animal models must be maintained in a heterozygous state. This means that only 25% of the offspring embryos are homozygous for the mutation of interest, and are by definition lethal. In *Drosophila*, such heterozygous mutants are maintained over highly rearranged chromosomes called "balancers", which prevent recombination (Miller et al., 2016, 2018). Here, together with my collaborators, we first devised an easy, streamlined and generalizable approach to profile single-cell genomic measurements from mutant embryos of any genotype (Figure 20). Rather than genotyping and hand sorting homozygous mutant embryos, I retained all embryos, which represent a pool of all three genotypes (homozygous loss-

of-function mutant, heterozygous, and homozygous wild-type) (Figure 20), and performed sci-ATAC-seq on their dissociated pooled nuclei. A technician in the lab (Rebecca Viales) sequenced the genetic background of the loss-of-function mutants while Mattia Forneris retrieved the sequence of the balancer chromosome (from (Ghavi-Helm et al., 2019)). Mattia Forneris then compared the sequence of the two genotypes (mutant background and balancer), calling SNPs and indels for each. Tobias Heinen then developed an algorithm to genotype each nucleus based on informative SNPs in the sci-ATAC reads. In contrast to standard allele imbalance studies, this requires relatively few informative reads per nucleus, as described below.

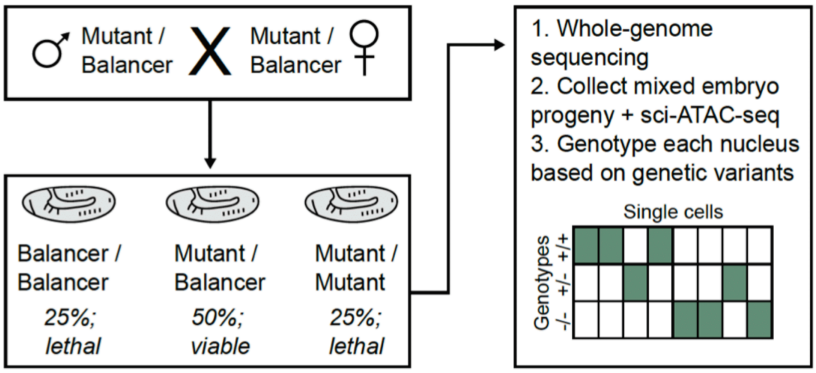

Figure 20. Illustration of the single-nucleus de-novo genotyping strategy.

Overview of the single-nucleus *de-novo* genotyping strategy. Pooled embryos of mixed genotypes are dissociated and their nuclei used for sci-ATAC-seq. Each nucleus is retrospectively *de-novo* genotyped based on discriminatory genetic variants in the ATAC-seq reads.

In all cases in this thesis where this strategy is applied, I performed the embryo collections and generated the sci-ATAC-seq data. Mattia Forneris computationally identified the discriminatory sequence variants needed to perform the genotyping. Tobias Heinen performed the computational assignment of genotypes to nuclei.

The characterized mutations of these mesoderm TFs were generated over twenty years ago and will have accumulated many additional mutations over the ensuing decades that could impact chromatin accessibility independently of the TF's function. To circumvent this, I first generated new loss-of-function alleles for each factor using CRISPR-Cas9 genome editing with single stranded oligo donors (ssODNs)(Gratz et

al., 2015) to recreate the characterized loss-of-function mutations for each factor in a common and fully sequenced isogeneic genetic background (for details please refer to Materials and Methods section 2.1: 'Generation of transcription factor mutants by CRISPR'). As the TFs are essential factors, these new CRISPR alleles are all homozygous lethal as expected, and, importantly, they non-complement the characterized loss-of-function allele when placed *in trans*, confirming that the lethality is due to the mutation of the TF and not a CRISPR off-target effect.

To apply the single-nucleus genotyping strategy (Figure 20), I collected staged embryos from the heterozygous adults (loss-of-function mutant^{CRISPR}/balancer chromosome), which were formaldehyde-fixed and processed for sci-ATAC-seq. The dissociated nuclei thereby come from a pool of F1 embryos with the following proportions and genotypes: 25% homozygous loss-of-function mutant/mutant, 50% heterozygous mutant/balancer, and 25% homozygous balancer/balancer. After sci-ATAC-seq, each nucleus was computationally genotyped *de novo* by Tobias Heinen based on the fraction of reads mapping to the mutant or balancer chromosomes, using a tool called Vireo (Huang et al., 2019). The genotype assignment is based on over 450,000 genetic variants between the balancer and the mutants common isogenic genetic background, which were computationally identified by Mattia Forneris. With a median of roughly 1,000 variants covered per cell, it was possible to genotype 99.9% of all theoretically assignable nuclei (Figure 21a) at very high confidence (> 0.9 posterior probability) (Figure 21b). The genotype assignment process is very robust, as 98% of assignments are identical when using sets of variants called at different quality thresholds (Figure 21c).

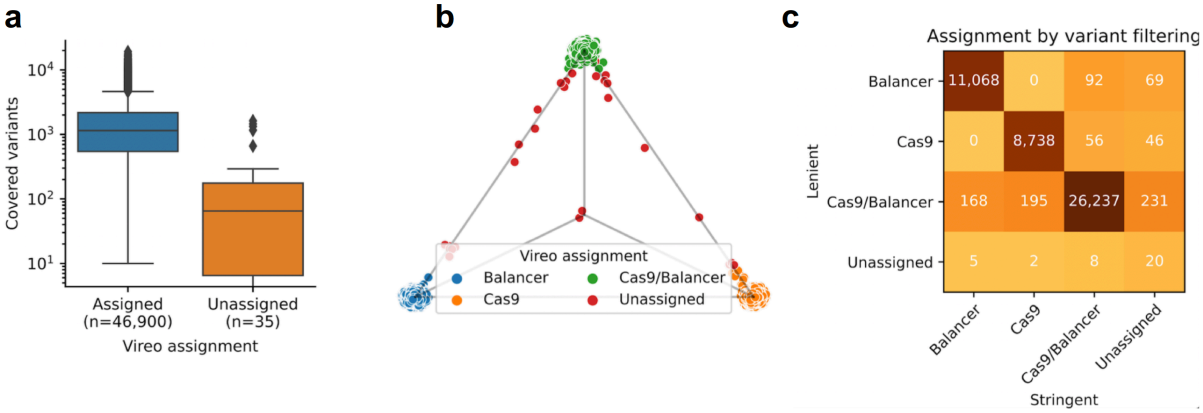

Figure 21. Quality control of genotype assignments.

(a) Box plots showing the total number of SNP variants per nuclei with a genotype assigned (blue) or unassigned (orange) by Vireo (Huang et al., 2019).
(b) Triangle plot showing assignment probabilities for each genotype, which are represented as points in the triangle. The center of the triangle corresponds to the vector (1/3, 1/3, 1/3), while the vertices represent a probability of 1 for Balancer (bottom left), Cas9 (bottom right) and Cas9/Balancer (top), respectively. Posterior probabilities are focused near the vertices of the triangle, indicating low posterior uncertainty.
(c) Confusion matrix for the genotype assignment using variants called with a stringent or a lenient filter, demonstrating that the model is robust to different reference annotations.
Mattia Forneris computationally identified the discriminatory variants needed to perform the genotyping.
Tobias Heinen performed the single-nucleus genotyping, analyzed the results and produced the plots for each panel of this figure.

The *de novo* single-nucleus genotyping approach has a number of advantages for single-cell profiling of mutants. It eliminates the need, and associated experimental time, to hand select embryos of the appropriate genotype and is therefore faster and more reliable. Profiling nuclei from homozygous and heterozygous siblings in the same experiment has an additional advantage for single-cell studies as it can aid in batch correction. As the heterozygous nuclei are essentially wild-type, they can be used to align mutant data from the same batch to the wild-type reference trajectory, thus avoiding 'over fitting' by batch aligners of biologically real mutant phenotypes. Moreover, having all loss-of-function mutations in the same genetic background allows the functional impact of different TF mutants to be directly compared.

4.9 Loss of the transcription factor Mef2 leads to a new cell state

Mef2 regulates differentiation of all major muscle types (Bour et al., 1995). To determine the functional impact of Mef2 on chromatin accessibility and mesodermal cell fate, I performed sci-ATAC-seq on *Mef2* mutant embryos (a pool of homozygous and heterozygous) at 10-12 hr, a time point that is crucial for muscle differentiation. The single-nucleus *de-novo* genotyping, performed by Tobias Heinen, assigned the

expected proportion of profiled nuclei as homozygous mutant (expected: 25%, observed: 26%) (Figure 22a). As these experiments were performed on whole embryos, I performed a first round of whole-embryo cell clustering to identify the muscle cells, which were then selected and re-clustered (2,567 muscle cells in total) (for details please refer to Materials and Methods section 4.10.1: 'Clustering of the *Mef2* mutant dataset'). This process resulted in five cell clusters with distinct chromatin accessibility, three of which could be identified as the somatic, cardiac and visceral muscle (Figure 22b).

Inspection of the proportion of homozygous mutant cells per cluster revealed that *Mef2* -/- mutant nuclei are almost entirely absent from the somatic muscle cluster (Figure 22c, d) and instead are highly enriched in two additional 'muscle clusters' with abnormal accessibility state, which appear close to, but distinct from, somatic muscle cells (Mutant1 and Mutant2 clusters). Mutant1 cluster is particularly enriched in *Mef2* -/- cells, representing 89% of the cluster's genotyped cells, while Mutant2 contains 56% of *Mef2* -/- cells (Figure 22c). This indicates that in the absence of *Mef2*, mesodermal cells are unable to establish the regulatory landscape necessary to become somatic muscle, and instead form a new altered state. Cells in the Mutant2 cluster have overall lower coverage (~2.2 fold lower) than the median coverage of other clusters. This may represent a cluster of more naive cells or cells undergoing apoptosis, although I cannot exclude that the clustering of these cells is driven primarily by their lower coverage.

To experimentally test the accuracy of the *de-novo* genotyping strategy, I hand sorted homozygous mutant embryos from the *Mef2* mutant based on a GFP marked balancer chromosome and I performed sci-ATAC-seq on these 100% *Mef2* homozygous nuclei. The hand sorted mutant nuclei show the same properties as the genotyped mutant nuclei (Figure 22c, d): they are absent from the somatic muscle cluster and accumulate in two mutant clusters – mainly in Mutant1 cluster, representing 88% of homozygous mutant cells for that cluster, similar to the computational genotyping above (Figure 22c). Both the digitally genotyped and hand-sorted homozygous mutant nuclei display the same alterations in chromatin accessibility at individual loci, as shown for two muscle contractile proteins *Mlc1* and *Msp300* (Figure 22e). Both genes have multiple Mef2-bound regions overlapping regions of open chromatin in cells from the somatic

cluster, which are almost completely closed in both the digitally genotyped and hand-sorted Mutant1 cells (Figure 22e). I also observe concordant gains in accessibility at regulatory regions for both the nuclear genotyped and hand-sorted homozygous *Mef2* mutants, as in the case of enhancer VT30021 (Figure 22e), which is embryonically active but normally not in muscle tissues. This proof-of-principle indicates that the *de-novo* nuclear genotyping strategy correctly assigns homozygous mutant nuclei.

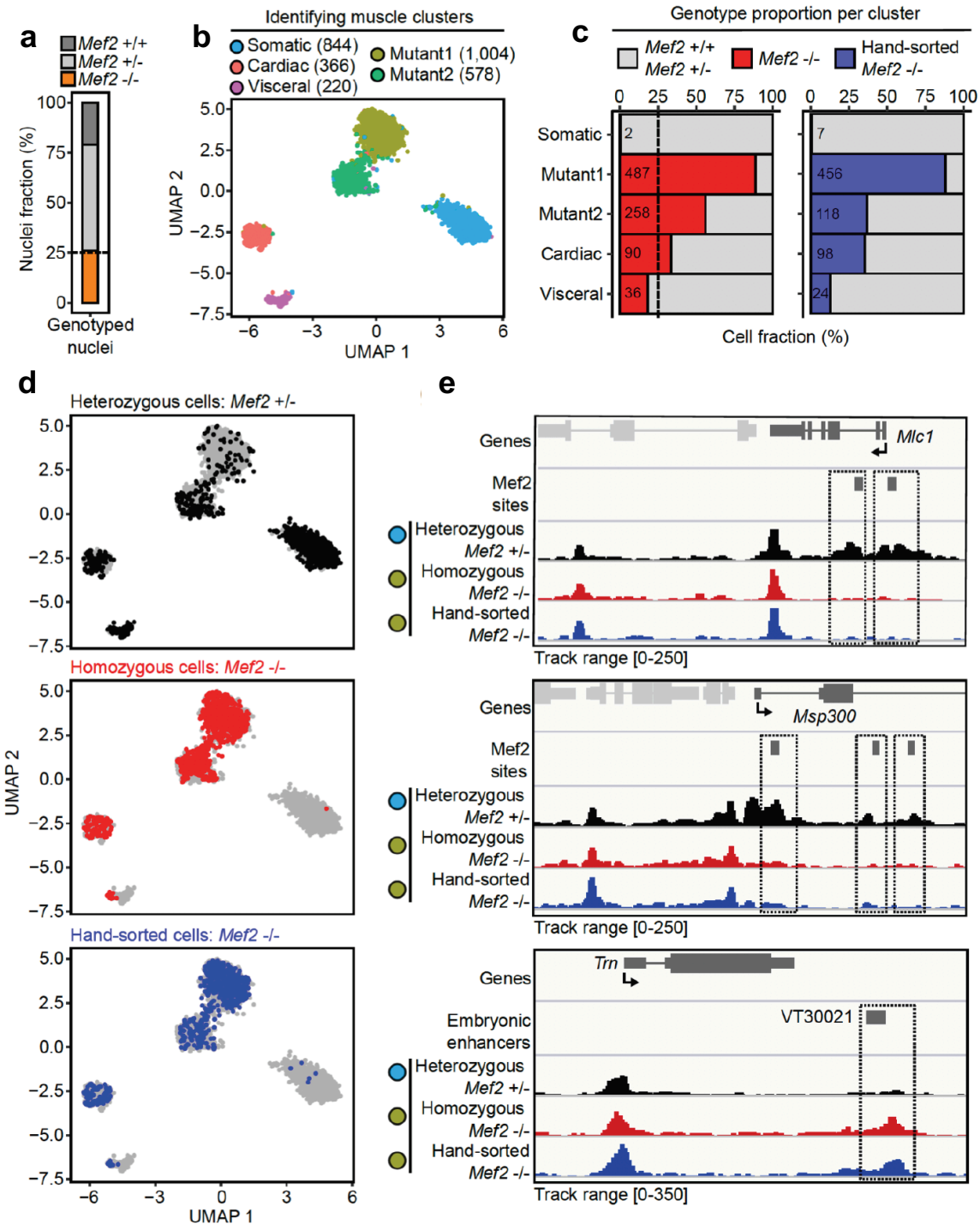

Figure 22. Loss of the transcription factor Mef2 leads to a new cell state.

(a) Proportion of nuclei assigned to each genotype. The expected proportion of *Mef2* homozygous (-/-) nuclei is 25%, as indicated by the dashed line.

(b) UMAP visualization of the clustering of muscle populations identified in the *Mef2* dataset, cells are colored by the inferred cell types.

(c) Proportion of genotypes in each cell cluster. Left: proportion of digitally genotyped homozygous mutant cells (*Mef2* -/-; red). Right: proportion of hand-sorted homozygous mutant cells (*Mef2* -/-; blue). The heterozygous (*Mef2*/balancer (+/-) and homozygous balancer cells are indicated in grey in both panels.

(d) Same as (b), heterozygous *Mef2* +/- cells are highlighted in black, digitally genotyped homozygous *Mef2* -/- cells in red and hand-sorted homozygous *Mef2* -/- cells in blue.

(e) Pseudobulk accessibility tracks for *Mef2* heterozygous (+/-) cells from the somatic cluster (black) and *Mef2* homozygous (-/-) cells digitally genotyped (red) or hand-sorted (blue) from Mutant1 cluster. Dot color indicates the cluster from (b).

In panels (a), (c), (d) and (e), the nuclei genotyping was performed by Tobias Heinen using a set of discriminatory variants identified by Mattia Forneris.

The whole-embryo single-cell data allowed me to explore if these *Mef2* mutant cells adopt another cell state, either from within the mesoderm or from any germ-layer. To assess this, I computed cluster-wise accessibility correlations of the Mutant clusters against all cell-types (both mesodermal and non-mesodermal cell clusters) in the embryo at 10-12 hr (Cusanovich et al., 2018a) (Figure 23). Both Mutant1 and Mutant2 are most highly correlated to clusters within the myogenic mesoderm, in particular the somatic muscle, and are clearly separated from the non-myogenic mesoderm, ectoderm and endoderm lineages (Figure 23), indicating that these cells are specified to become muscle, but may become blocked in their development.

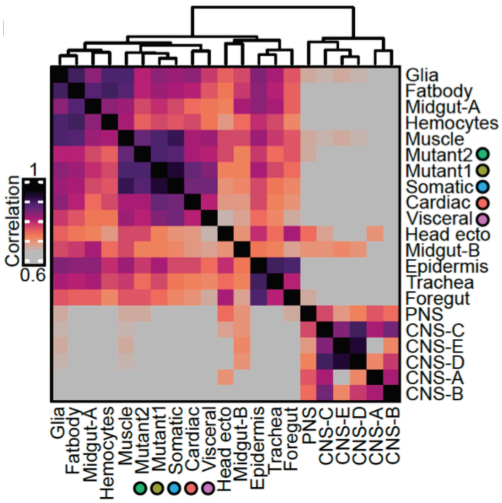

Figure 23. Chromatin accessibility correlation across embryonic populations and *Mef2* mutants.

Pearson correlation matrix of cluster-wise accessibility for muscle and Mutant clusters identified in Figure 22b and diverse cell clusters from whole embryos, identified in (Cusanovich et al., 2018a). Dot color indicates the cluster from Figure 22b.

To determine if the mutant cells are stuck in an earlier myogenic state, I combined the mutant data (combining the digitally genotyped and hand-sorted mutant cells, given that they appear identical in nature) and the heterozygous cells with all cells in the wild-type reference trajectory (Figure 13a) and re-clustered the data (Figure 24). The heterozygous cells (the wild-type clusters in Figure 22b) behave indistinguishably from the reference cells, falling within the expected wild-type populations on the trajectory. In contrast, *Mef2* -/- Mutant1 and Mutant2 cells cluster separately, off the wild-type muscle trajectory, but roughly at the appropriate 'temporal' time-point (Figure 24). If these cells were blocked in their developmental progression, I would have expected them to cluster on the trajectory at some earlier time-point in development, which is not the case.

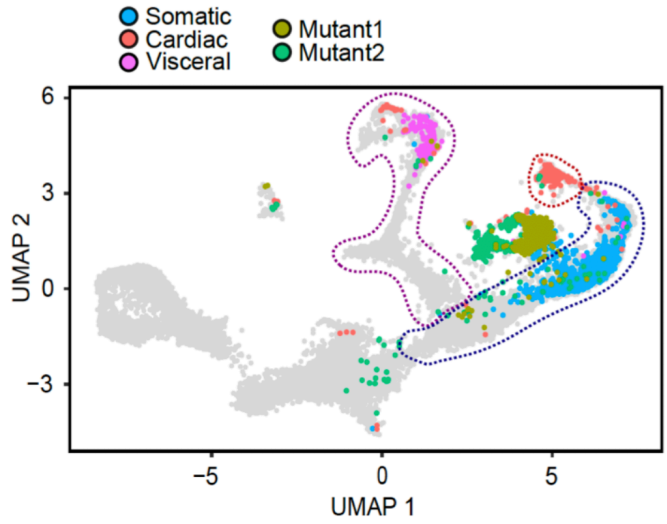

Figure 24. Co-clustering the *Mef2* mutant dataset with the wild-type reference trajectory.

UMAP visualization of the muscle populations identified in Figure 22b, including *Mef2* +/- and -/- cells, co-clustered with the wild-type muscle time course from Figure 13a. Cell clusters identified in Figure 22b are plotted on top of the wild-type time course (cells colored in grey). The somatic, cardiac and visceral populations from the wild-type time course are enclosed by blue, red and purple dashed lines, respectively. Note the *Mef2* -/- clusters 1 and 2 are located close to the somatic muscle (blue) but off the wild-type trajectory.

To explore this further, I computed cluster-wise accessibility correlations of the Mutant and muscle clusters for each time point, which confirmed that both Mutant1 and Mutant2 are progressing to the appropriate developmental stage (Figure 25). These cells are therefore not simply immature muscle cells, but have rather developed a new abnormal 'muscle-like' state at roughly the correct corresponding stage of the wild-type trajectory. Taken together, this suggests that *Mef2* is not only required as a differentiation factor to regulate the expression of muscle contractile genes, but also to prevent cells from undergoing other cell state changes.

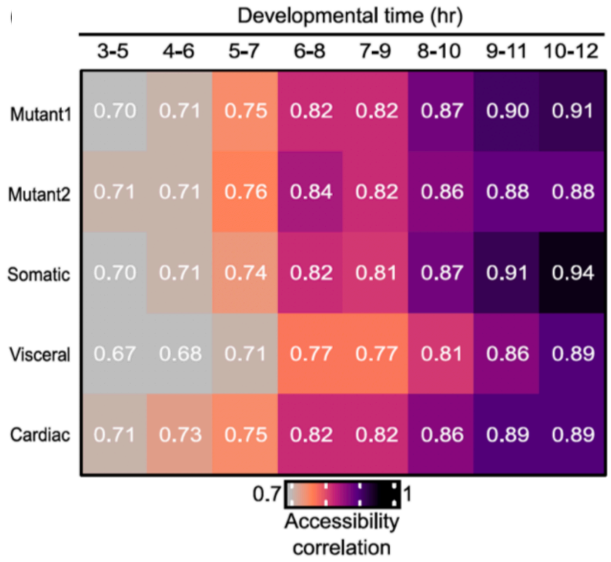

Figure 25. Temporal correlation between the *Mef2* dataset clusters and the wild-type reference trajectory.

Pearson correlation of chromatin accessibility for the cell clusters identified in Figure 22b against time point aggregate samples from the wild-type muscle time course. Mutant 1 cells have the highest correlation to cells at 10-12hr, matching the specific time point of their sampling (10-12 hr).

4.10 Loss of *tinman*, *bagpipe* and *biniou* differentially alters cellular composition

I applied the same approach to three other loss-of-function mutants for TFs involved in the specification of the dorsal mesoderm (*tinman*), and its derived visceral muscle (*bagpipe*, *biniou*) that forms the gut musculature (Azpiazu and Frasch, 1993; Zaffran et al., 2001). These TFs have a hierarchical relationship between them, where Tinman regulates Bagpipe expression at stage 10 (6-8 hrs of development), which in turn regulates Biniou expression (Figure 26). For all three mutants, I performed sci-ATAC-seq on a pool of homozygous and heterozygous embryos, as above. Staged *bagpipe* and *biniou* mutant embryos were collected at 6-8hr and the mesodermal population isolated by Mef2 FAC sorting, as in the wild-type trajectory, obtaining high quality profiles for 6,306 and 5,833 mesodermal cells in these mutant embryos, respectively. This pre-sorting for the mesodermal population could not be performed for *tinman*, as it regulates *Mef2* expression. I therefore performed sci-ATAC-seq on whole embryos

of *tinman* mutants, at 5-7 hr, and then performed a first round of clustering to identify 6,786 high-quality mesodermal cells.

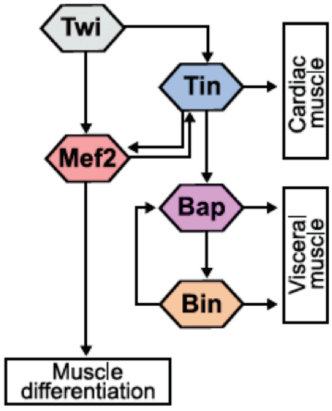

Figure 26. Illustration of the mesoderm TF network.

Simplified schematic of the mesoderm regulatory network dissected from genetic studies. Adapted from (Zinzen et al., 2009) with permission License Number: 5118140343137.

To assess the fate of the mutant cells, I directly compared their development to the wild-type trajectory by co-clustering the combined mutant data, representing 18,925 homozygous and heterozygous cells, together with the wild-type mesoderm time course, correcting for batch-level effects with Harmony (Korsunsky et al., 2019) (for details please refer to Materials and Methods section 4.10.2: 'Clustering of *tinman*, *bagpipe* and *biniou* mutant datasets'). Re-clustering and re-annotation of this joint dataset, representing 40,232 cells, revealed a structure that is generally consistent with the clustering of the wild-type dataset alone and that reveals a similar cell-type composition (Figure 27).

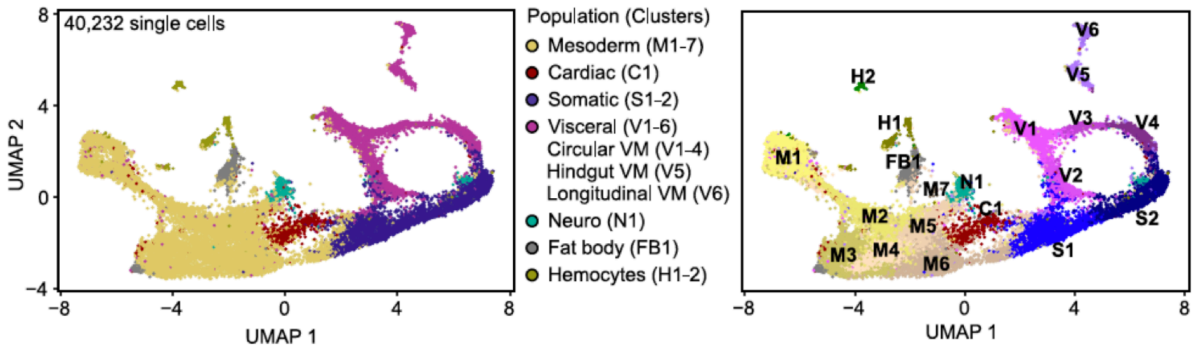

Figure 27. Co-clustering the *tinman*, *bagpipe*, *biniou* datasets with the wild-type mesoderm reference.

UMAP visualization of the wild-type sci-ATAC-seq time course re-clustered with single cell data from *tinman* (*tin*), *bagpipe* (*bap*) and *biniou* (*bin*) heterozygous (+/-) and homozygous (-/-) embryos. Each dot represents a single cell, colored by cell-type annotation (left) and by cell cluster (right).

Nuclei from heterozygous mutant cells (+/-) for *tinman* at 5-7 hr are present in the cardiac (cluster C1) and early visceral muscle clusters (clusters V1-3, plus V5 and V6) (Figure 28a, middle panel, black). Similarly, *bagpipe* and *biniou* heterozygous mutant cells are present in the visceral mesoderm, spanning both branches, and extending to later stages of embryogenesis (clusters V1-4, plus V5 and V6) (Figure 28b, c, middle panel, black). Examining the homozygous mutant nuclei (-/-) revealed that, in contrast to *Mef2*, the proportion was significantly lower than the expected 25% for all three mutants, representing 15%, 19% and 17% for *tinman*, *bagpipe* and *biniou*, respectively (Figure 28a, b, c, left). This indicates that a proportion of homozygous mutant cells are not maintained, suggesting that they likely undergo apoptosis as they cannot progress in their development. The trajectories of the remaining mutant cells (-/-) are very different from their heterozygous siblings. *tinman* homozygous mutant cells are completely absent from the cardiac lineage, with few remaining cells in the early visceral muscle clusters (Figure 28a, middle panel, red, V1, V2) and almost no late stage visceral muscle cells (V3-V6). Interestingly, the hindgut and longitudinal visceral muscles appear largely unperturbed by *tinman* loss-of-function (clusters V5 and V6, Figure 28a). Moreover, there is a significant reduction of homozygous mutant cells in late mesoderm stages (clusters M4, M5), which may represent the dorsal mesoderm. Similarly, the *bagpipe* and *biniou* homozygous mutant nuclei are absent from clusters at later stages of visceral muscle development (clusters V2 and V3, Figure 28b, c middle panel, red). The early visceral cells (cluster V1) are more prominently affected in *tinman* mutants, and to a lesser extent in *bagpipe* and *biniou* mutants, reflecting the hierarchical position of these TFs with Tinman acting upstream of both factors. Other visceral muscle subtypes (hindgut visceral muscle (cluster V5) and longitudinal visceral muscle (cluster V6) are not significantly affected, as previously suggested but not shown (Zaffran et al., 2001). These findings indicate that the circular VM cells are initially specified in *bagpipe* and *biniou* mutant embryos, but are blocked from further

expansion and differentiation, resulting in a loss of the VM at later stages (clusters V2 and V3).

This molecular data therefore mirrors the high-level phenotypes described for *tinman*, *bagpipe* and *biniou* mutants as seen by immuno-staining of mutant embryos (Azpiazu and Frasch, 1993; Zaffran et al., 2001). However, this single-cell approach also reveals more fine-grained phenotypes that were not previously observed, including a gain of mutant cells in other muscle lineages, suggesting a change in cell state in these mutant embryos. For example, there is a significant over-representation of *tinman* mutant cells in the early mesoderm (cluster M2) and in the somatic lineage (cluster S1) (Figure 28a, right). The removal of these TFs thereby not only results in a loss of tissue (one cell fate), but also a more subtle gain of cells dispersed in other tissues from different mesodermal trajectories, highlighting the plasticity of cell fates within the myogenic mesoderm.

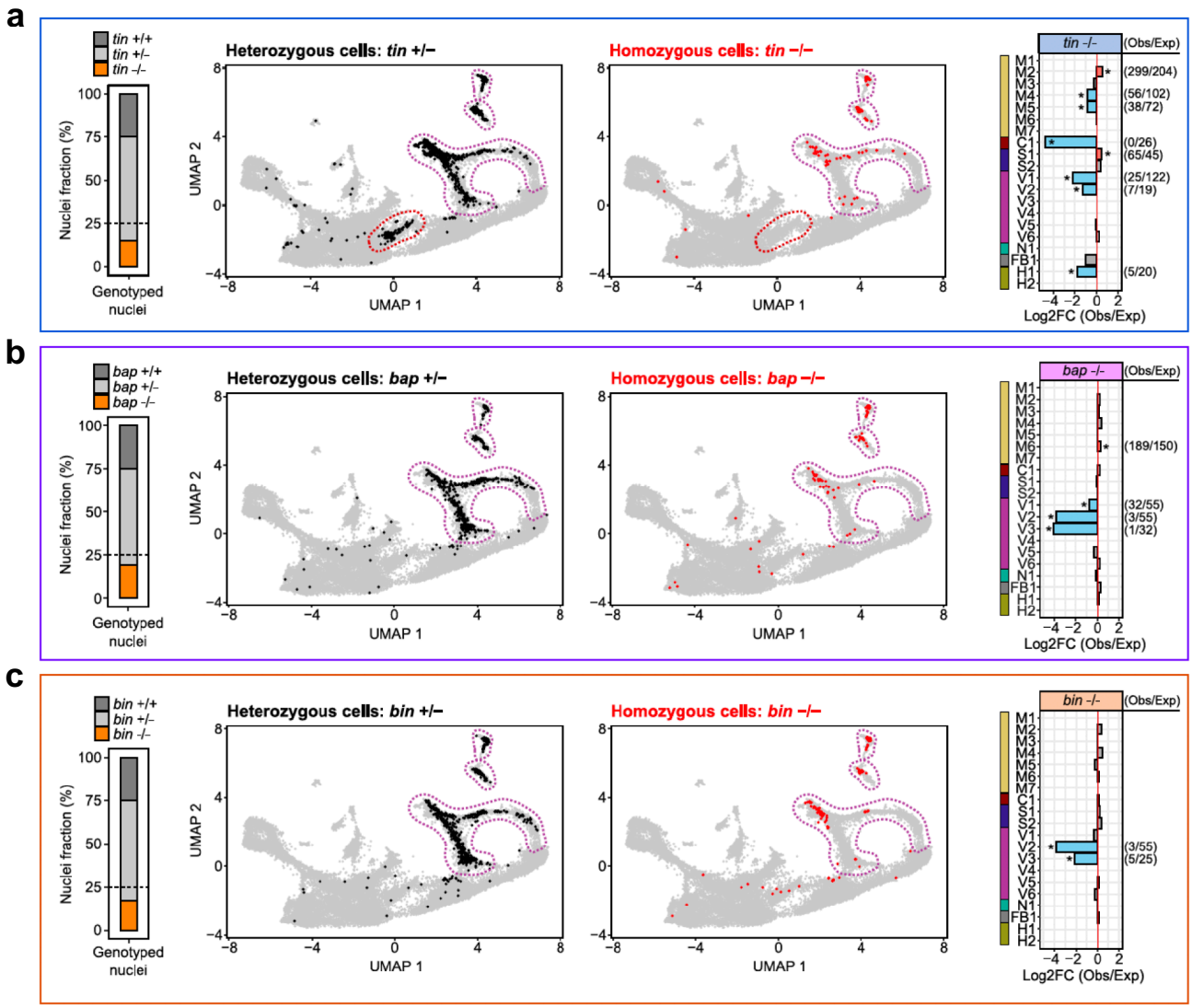

Figure 28. Loss of *tinman*, *bagpipe* and *biniou* differentially alter cellular composition.

(a) Left: proportion of nuclei assigned to each genotype. The expected proportion of homozygous -/- mutant nuclei is 25%, indicated by the dashed line. Middle: same as Figure 27, where cells from *tin* embryos (5-7 hr) are colored by genotype (heterozygous *tin* +/- cells in black; homozygous *tin* -/- cells in red) and the cardiac (dark red in Figure 27) and visceral (purple in Figure 27) muscle populations are highlighted with dashed lines. Only cells belonging to the cardiac and visceral populations are plotted. Right: Log2 foldchange of observed versus expected homozygous *tin* -/- cells in each cell cluster (shown for clusters with >100 cells). Asterisks indicate clusters with a significant (Fisher's exact test (P-value < 0.05)) over- (red bars) or under- (light blue bars) representation of *tin* -/- cells, the number of observed/expected cells is indicated (right).

(b) Same as (a), highlighting cells from *bagpipe* heterozygous and homozygous embryos (6-8 hr): *bap* +/- in black; *bap* -/- cells in red.

(c) Same as (a), highlighting cells from *biniou* heterozygous and homozygous embryos (6-8 hr): *bin* +/- in black; *bin* -/- cells in red.
In panels (a), (b), and (c), the nuclei genotyping was performed by Tobias Heinen using a set of discriminatory variants identified by Mattia Forneris.

4.11 Removal of Mef2 affects chromatin accessibility at hundreds of regulatory regions

As this is the first time, to my knowledge, that single cell ATAC-seq has been applied to TF mutants in the context of a developing embryo, it provides a unique opportunity to explore the extent to which such single cell data can discern regulatory properties of the TF or its enhancers. As a proof of principle, I focused on the *Mef2*-/- mutant cell cluster (Mutant1), as it contains the highest number of mutant cells (943 cells). The somatic muscle cluster is the closet cell-type to Mutant 1 (Figure 23). Of the 8,725 accessible regions in both cell clusters (Mutant1 and the somatic muscle cluster), 408 have significant differential accessibility (DA) in *Mef2*-/- mutant cells (log2 fold change > +/- 0.5, Bonferroni corrected p-value < 0.05) (Figure 29, Data Table 2). The majority of DA sites have reduced, compared to gained, accessibility (274 reduction, 134 gain) and reduced sites often have a larger fold-change (Figure 29).

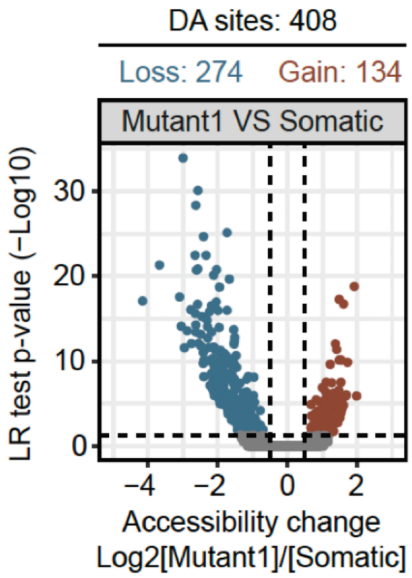

Figure 29. Differential accessibility analysis of *Mef2* mutant versus somatic muscle.

Volcano plot of 408 differentially accessible (DA) sites. X-axis indicates log2 fold-change in chromatin accessibility, showing an increase (positive values) or decrease (negative values) in Mutant1 compared to the somatic muscle cluster. Y-axis indicates the Bonferroni corrected P-value from the logistic regression test used to identify the DA sites (-Log10 scale).

To explore the 408 DA sites further, I first categorized them into Mef2-bound and unbound sites, using bulk Mef2 ChIP data at multiple time-points of embryonic development from our lab (retrieved from (Zinzen et al., 2009)). Almost half of the DA regions (48%,197/408) are bound by Mef2 at this stage or earlier in embryogenesis. Mef2-bound DA sites almost exclusively lose accessibility (Figure 30), comprising 66% of all DA sites with reduced accessibility. In contrast, regions that gain accessibility are generally not bound by Mef2 (Figure 30).

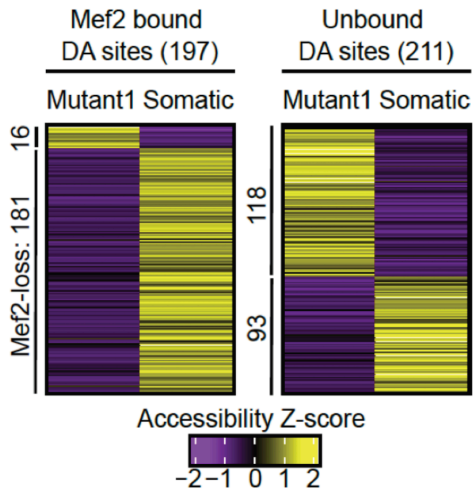

Figure 30. Heatmaps of DA sites split by Mef2 occupancy.

Heatmaps of chromatin accessibility at the 408 DA sites, split by those bound by Mef2 during embryogenesis (Mef2-bound; 197 DA sites) or not-bound (Unbound; 211 DA sites). Over 90% (181/197) of regions bound by Mef2 have reduced accessible in *Mef2* mutant cells (Mutant1).

4.12 Mef2 is required for chromatin accessibility at its high affinity sites

Removal of Mef2 affects the accessibility of only a specific subset of Mef2-bound sites, which represent ~15% of all Mef2-bound regions. This raises the question, what distinguishes Mef2-bound sites that are sensitive to, or resistant to, Mef2 removal?

This could depend on how extensively the region is occupied by other factors (i.e. its degree of combinatorial binding) or on the affinity of Mef2 binding. To distinguish between these two possibilities, I first examined the co-occupancy of ten mesodermal TFs at Mef2-bound regions. Susceptible sites are generally less frequently occupied by other mesodermal TFs compared to non-susceptible sites (Figure 31a). Furthermore, the fraction of DA sites tends to decrease as sites are bound by an increasing number of mesodermal TFs (Figure 31b). Examining the occupancy of a much larger number of TFs, 280 factors from the modERN consortium (Kudron et al., 2018), also shows an inverse correlation between differential chromatin accessibility in *Mef2*-/- cells and the number of bound TFs (Figure 31c): the median number of bound TFs is 3 for the DA class and 40 for non-DA sites (Figure 31d). Therefore, sites that require Mef2 for their accessibility tend to be bound by Mef2 alone or with a small number of factors, perhaps cooperatively, suggesting that these regions are very Mef2 dependent.

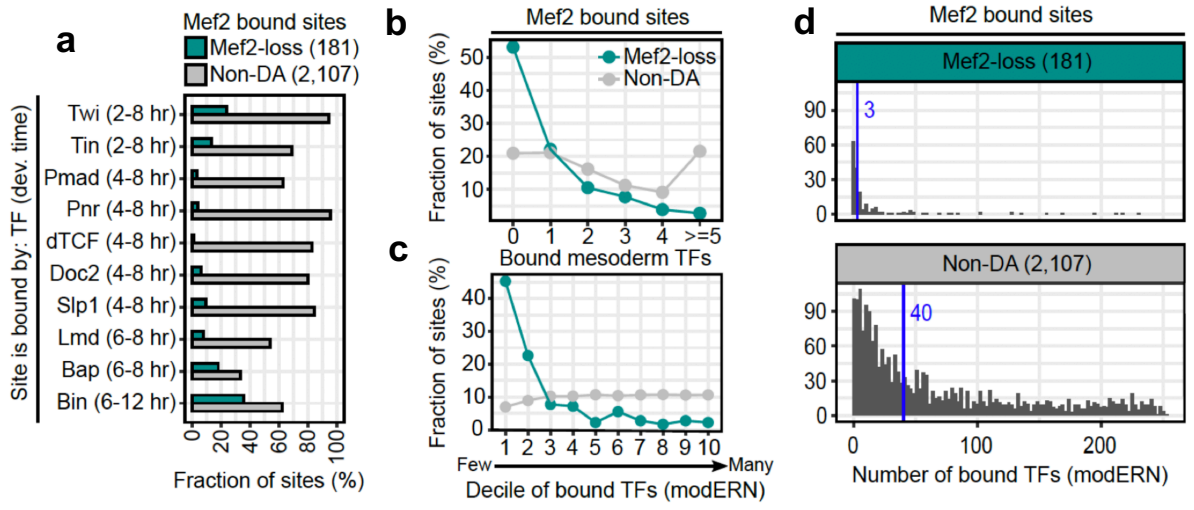

Figure 31. Co-occupancy of Mef2 and other TFs at Mef2-bound DA and non-DA sites.

(a) Fraction of Mef2-bound sites occupied by additional mesodermal transcription factors for DA sites with reduced accessibility (Mef2-loss; green) or with unchanged accessibility (non-DA; grey) in *Mef2* mutant cells. TF occupancy was measured by ChIP for each factor at the embryonic time windows indicated.

(b) Fraction of Mef2-bound sites occupied by the indicated number of mesodermal TFs for DA sites with reduced accessibility (Mef2-loss; green) and non-DA sites in *Mef2* mutant cells. X-axis value 0 indicates sites bound by Mef2 and no additional mesodermal TF.

(c) Fraction of Mef2-bound sites occupied by an increasing number of general TFs (retrieved from the modERN database (Kudron et al., 2018)) stratified by decile for DA Mef2-loss and non-DA sites.

(d) Number of general TFs (retrieved from the modERN database (Kudron et al., 2018)) that occupy DA Mef2-loss and non-DA Mef2-bound sites. Blue line indicates the median number of bound TFs.

Occupancy data for (a) and (b) was retrieved from datasets of mesodermal TFs produced by our lab (Cunha et al., 2010; Jakobsen et al., 2007; Junion et al., 2012; Zinzen et al., 2009).

Occupancy data for (c) and (d) was retrieved from (Kudron et al., 2018).

To investigate this further, I used the quantitative Mef2 ChIP signal (peak height) as a proxy for Mef2 affinity to different sites. Susceptible (DA) sites have significantly higher Mef2 ChIP signal compared to non-susceptible (non-DA) sites (Figure 32a). Plus, the proportion of DA sites steadily increases as the Mef2 ChIP signal increases, going from 5% of DA sites in the lowest to 35% in the highest ChIP quantile (Figure 32b). This indicates that sites bound more strongly by Mef2 are more likely to have reduced chromatin accessibility upon Mef2 removal. In support of this, although both classes are occupied by Mef2, susceptible sites have a 2.5 fold-enrichment (61% DA group vs 17% non-DA group, Fisher's exact test p-value = $2x10^{-25}$) in the presence of the Mef2 motif (Figure 32c). These results indicate that Mef2 is required to establish and/or maintain chromatin accessibility at a large fraction of its high affinity sites.

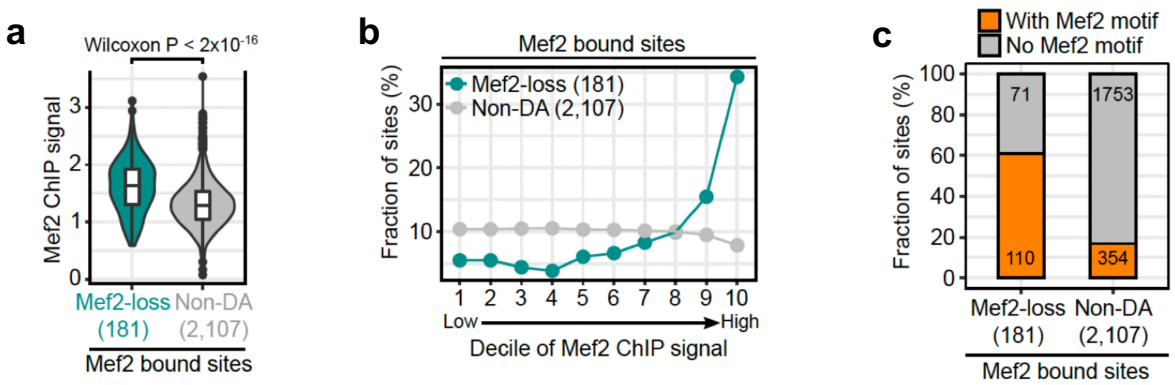

Figure 32. Mef2 is required for chromatin accessibility at its high affinity sites.

(a) Quantification of Mef2 ChIP signal at DA Mef2-loss and non-DA Mef2-bound sites. Wilcoxon P-value indicated.

(b) Fraction of Mef2-bound sites stratified by decile of increasing Mef2 ChIP signal, shown for both DA Mef2-loss (green) and non-DA sites (grey).

(c) Fraction of DA Mef2-loss and non-DA Mef2-bound sites with (orange) or without (grey) a Mef2 motif.

Mef2 occupancy data was retrieved from (Zinzen et al., 2009).

4.13 Mef2 is required for proper expression of key muscle genes

Many of the DA Mef2-loss regions overlap characterized (Figure 33a) or putative (Figure 33b) muscle enhancers. The loss of accessibility at these sites may therefore lead to changes in the expression of mesoderm/muscle genes (Figure 33c), which most likely contributes to the *Mef2* mutant phenotype.

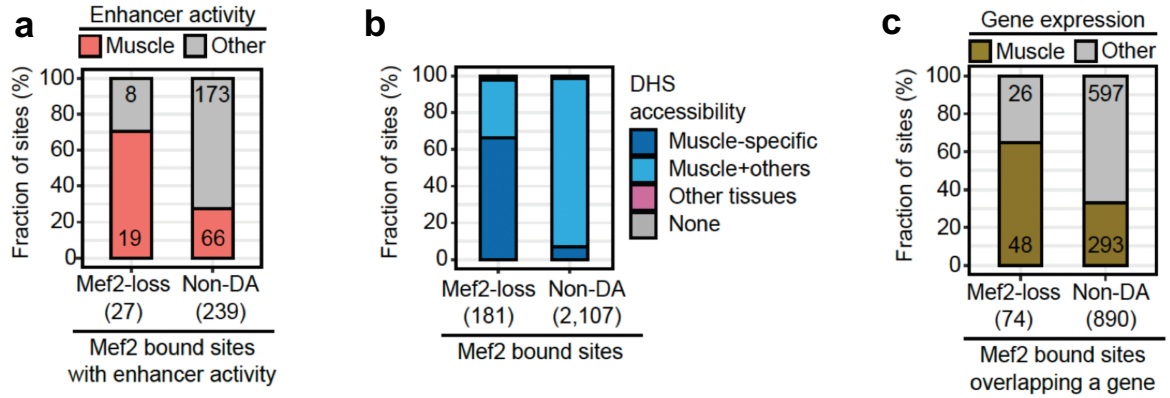

Figure 33. DA Mef2-loss sites frequently overlap muscle enhancers and genes.

(a) Fraction of DA Mef2-loss and non-DA Mef2-bound sites that overlap characterized embryonic enhancers with demonstrated activity in muscle (red) or other tissues (grey).

(b) Fraction of DA Mef2-loss and non-DA Mef2-bound sites that overlap DNase I Hypersensitive Sites (DHS) accessible in muscle only (Muscle-specific), in muscle and other tissues (Muscle+others), in other tissues only (Other tissues) or that do not overlap a DHS (None). DHS data was retrieved from (Reddington et al., 2020).

(c) Fraction of DA Mef2-loss and non-DA Mef2-bound sites that overlap genes with expression in muscle (brown) or other tissues (grey).

To examine this in more detail, I integrated bulk expression data from *Mef2* mutant embryos (Sandmann et al., 2006) and looked for genes with a Mef2-bound site within their vicinity (defined as 5kb upstream and in intronic regions). Using this metric, 1,705 genes are associated with at least one Mef2-bound open chromatin region. Differentially expressed (DE) genes (log2 fold change > +/- 0.7, q <0.05) associated with regions with reduced accessibility in *Mef2* -/- mutant cells are highly over represented among genes with downregulated, but not upregulated (30% vs 8%), expression in *Mef2* -/- mutants (Figure 34a). Moreover, genes associated with reduced Mef2-bound sites have significantly stronger changes in both their chromatin accessibility (Wilcoxon p-value = $4x10^{-13}$) and gene expression (Wilcoxon p-value =$9x10^{-4}$), compared to genes with unchanged Mef2-bound sites (Figure 34b). Many Mef2 known targets genes are among this set, including *Mhc*, *Mlc1/2*, *Tm1*, *Mp20*, *Mlp60A and Msp300*. In addition, their expression changes become more severe with increasing numbers of associated regulatory regions with reduced accessibility (Figure 34c). These findings indicate that Mef2 functions primarily as an activator, and as the predominant regulator for the expression of these genes, which in turn likely leads to the muscle differentiation defects observed in *Mef2* mutant embryos. It is also a rare example demonstrating that a single TF can have a cumulative effect on a gene's expression through the action of multiple independent enhancers. In Mef2's case, it does this by affecting the regulation of many genes in a similar manner.

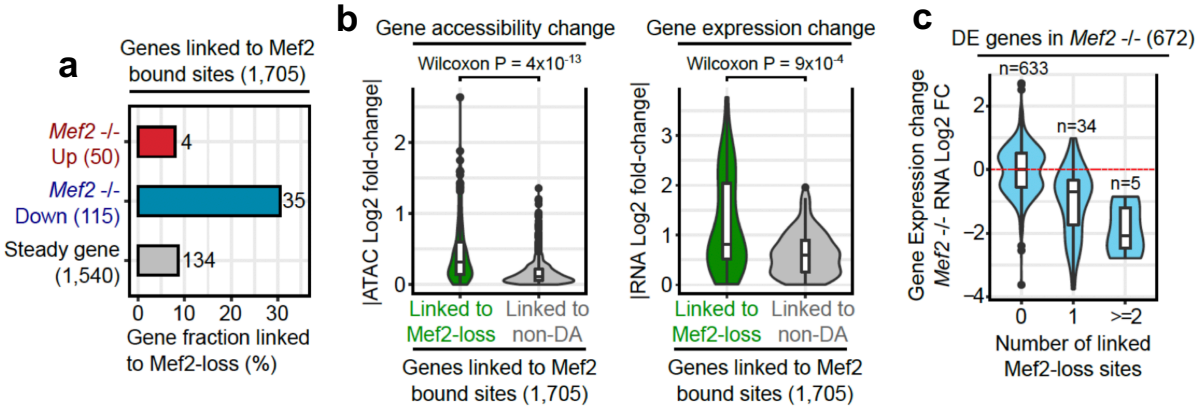

Figure 34. Loss of Mef2 frequently affects muscle genes expression.

(a) Fraction of Fraction of genes linked to a DA Mef2-loss site split by gene expression state (up, down or unchanged) in *Mef2* -/- embryos.

(b) Absolute ATAC (left) and RNA (right) Log2 fold-change in *Mef2* -/- embryos, split by genes linked to DA Mef2-loss (green) or non-DA (grey) Mef2 sites.

(c) RNA Log2 fold-change of differentially expressed genes in *Mef2* -/- embryos associated with an increasing number of DA Mef2-loss sites.

The gene expression data in *Mef2* -/- embryos used in each panel of this figure was retrieved from (Sandmann et al., 2006).

4.14 On-going experimental validations

The analysis presented in section 4.6 indicates that the cumulative chromatin accessibility level across the body of a gene can be used to approximate the gene's expression. Of the 864 genes identified with differential accessibility across lineages (Figure 17a), many are known TFs that regulate muscle sub-types while many others are completely uncharacterized. To discover new TFs that might regulate muscle development, I selected six uncharacterized TFs for further investigation. First, I plan to perform in-situ hybridizations for these genes, to verify that they are expressed in the muscle population predicted by my analysis. Second, I plan to knock them down by RNA interference, using available fly lines, and perform immunostaining for common muscle markers to determine the impact on muscle development.

The analysis presented in section 4.11 revealed that hundreds of regulatory regions, including putative enhancers, are affected in *Mef2* mutant embryos (for example in Figure 22e). The changes in chromatin accessibility at these sites could reflect a change in activity of the underlaying enhancers. From the list of differentially accessible regions in *Mef2* mutants (Data Table 2), I searched for sites that overlap *in vivo* validated embryonic enhancers and for which reporter fly lines are available in the Vienna and Janelia stock collections. Among these, I selected six cases in which my sci-ATAC-seq data predicts a loss of muscle enhancer activity in *Mef2 -/-* embryos and three cases in which it predicts that the enhancer will gain muscle activity in the *Mef2*

mutants. The reporter fly lines carry a construct with the enhancer element driving the expression of a reporter gene. These enhancer-reporter lines are currently being crossed into a *Mef2* loss-of-function mutant background. I will then amplify the lines, collect embryos and perform *in-situ* hybridization against the reporter gene, to validate the predicted changes in enhancer activity in *Mef2* mutant embryos.

4.15 Discussion and Conclusions

In this chapter, I applied sci-ATAC-seq to generate a rich atlas of chromatin accessibility for over 20,000 cells at high temporal resolution during *Drosophila* mesoderm development in wild-type embryos. With this approach I was able to capture the specification of all myogenic lineages from an entire germ-layer and chart its continuous progression towards diverse cellular fates. Chromatin accessibility was sufficient to resolve all the major muscle populations, the somatic, the visceral and cardiac muscle, and it even allowed to identify and the chart the progression of heterogenous sub-populations within the visceral muscle. Embryonic muscle lineages are therefore characterized by very distinct and dynamic regulatory landscapes.

Profiling development at high temporal resolution over a dense time course brings several advantages. I demonstrated that this approach can refine the approximate tissue and temporal window of TF activity, even when occupancy data was profiled in bulk at low temporal and spatial resolution. My dataset can therefore be used to revisit collections of bulk ChIP data to reveal the cellular context in which TFs are active. Moreover, I demonstrated that, when profiled in such a dense manner, chromatin accessibility data is sufficient to reconstruct continuous developmental trajectories and track the usage of regulatory elements along lineage progression. The data itself provides a rich resource of regulatory changes associated with each step of mesoderm specification into different muscle types, which can serve as a basis for the community to discover new principles of gene regulation in future studies. To facilitate access to the data, I set up a website that enables the community to interactively explore the UMAP plots and also download data objects for further analysis: http://furlonglab.embl.de/ss/Drosophila-Mesoderm-Chromatin-Accessibility/.

In this chapter I also presented a general framework for dissecting TF function at both a cellular and molecular level using a systematic, unbiased single-cell profiling approach. In classic genetic studies, phenotypes of developmental mutants are typically assessed by immunostaining with markers for different tissues, and often described in qualitative and somewhat arbitrary terms. In many cases phenotypes were missed in the original assessment, because the tissue was outside the scope of the study or because suitable tissue markers were not available. In addition, translating high-level phenotypes, such as tissue defects, to the underlying molecular function of the TF remains a challenge, and typically the regulatory input is only assessed by occupancy in wild-type embryos compared to gene expression changes in the mutant.

In this chapter I showed how single-cell regulatory trajectories, obtained by a dense time-course of wild-type embryos development, provides a new opportunity to map developmental mutants to more precise cell states, thereby providing more fine-grained insights into mutant phenotypes. In the four mutants I assessed, this approach not only revealed the loss of the expected cell types, as previously reported, but also the ability of some mutant cells to join seemingly normal trajectories of other muscle lineages. Going forward, this approach could be applied to reassess phenotypes and regulatory programmes of ‘classic’ developmental mutants, and also to uncover phenotypes of completely uncharacterized mutants *de novo*, thus contributing to the construction of predictive networks of tissue and organismal development.

4.16 Contributions

I performed all the experiments and computational analysis presented in this chapter, unless otherwise stated.

Dr. Mattia Forneris (*Bioinformatician, Furlong lab*) identified the discriminatory genetic variants used for the single-nucleus *de-novo* genotyping.

Tobias Heinen (*PhD student, Stegle lab*) computationally assigned the genotypes to nuclei and generated the plots presented in Figure 21.

Dr. Rebecca Rodriguez Viales (*Technician, Furlong lab*) generated the whole-genome sequencing (WGS) library of the isogenic vasa-Cas9 line (the mutant background).

Dr. Alessandra Reversi of the EMBL *Drosophila* injection service performed the embryo injections to generate the CRISPR mutant lines.

Staff at the EMBL Genecore Facility carried out all sequencing. I prepared all the libraries myself and submitted them to the facility for sequencing. The only exception was the vasa-Cas9 WGS library, which was prepared and submitted by Dr. Rebecca Rodriguez Viales, as indicated above.

Dr. Oliver Stegle (*Group Leader, Stegle lab*) supervised the work of Tobias Heinen.

Dr. Eileen Furlong (*Group Leader, Furlong lab*) supervised the project, contributed to the ideas and discussions and provided funding.

5 Using sci-ATAC-seq to uncover cell-type specific genetic perturbations

Enhancers tightly control the spatiotemporal expression of genes driving embryonic development. By displaying multiple binding sites for different TFs, which bind combinatorially to modulate gene expression, enhancers can integrate complex developmental signals. Understanding how regulatory information is encoded in the enhancer sequence is crucial for dissecting their function role in regulatory networks. In the previous chapter (Chapter 4) I demonstrated an approach by which enhancers can be perturbed in *trans* by mutating key developmental TFs. On the other hand, introducing large-scale perturbations in *cis,* by mutating enhancer sequences, is less straightforward, especially in embryos. Methods such as Perturb-ATAC (Rubin et al., 2019) and Spear-ATAC (Pierce et al., 2021), which combine CRISPR screening with single-cell regulatory read-outs, are difficult to extend beyond cell culture models, because our capacity to mutagenize multiple loci in higher organisms and handle the husbandry of large numbers of mutant lines is still limited.

In this regard, naturally occurring genetic mutations within regulatory elements can be considered as an alternative means of perturbations in *cis*. Individuals of the same species carry a large degree of natural sequence variation in their DNA, in the form of single-nucleotide polymorphisms (SNPs), small deletions (indels) and other genetic alterations, which can all lead to phenotypic variation. Genetic variants have been associated with alterations in molecular phenotypes, including chromatin accessibility (Behera et al., 2018; Degner et al., 2012; Floc'hlay et al., 2021), histone modifications (Floc'hlay et al., 2021; Sun et al., 2016), TF binding (Behera et al., 2018; Kasowski et al., 2010), gene expression (Cannavò et al., 2017; Floc'hlay et al., 2021), and high-level phenotypes, most notably a wide range of human diseases (Manolio, 2010). Genome-wide profiling revealed that most genetic variation occurs within the non-coding genome (Manolio, 2010) and thus it can affect enhancer function, for example by breaking TF binding sites. A well-known case is that of human variant rs11708067, which disrupts the function of the enhancer it resides in, leading to lower expression of its target gene, *ADCY5*, and in turn a higher risk of developing diabetes (Roman et al., 2017). However, in the majority of cases, the mechanism by which genetic variants

lead to a phenotypical alteration is not known. This lack of functional knowledge particularly limits our ability to predict the effect of genetic variants.

Unraveling the genotype-to-phenotype relationship cannot take place without first understanding the cellular context in which genetic variants have an impact. The effect of genetic variants can be homogenous and pervasive across most cell-types in an organism. But, as enhancers often act in a tissue-type specific manner, genetic variation within enhancers can differentially impact diverse cell-types and cellular states. Thus, genetic variants can also have a heterogenous context-specific effect. Due to technical challenges, most studies so far have been performed using bulk sequencing methods on very heterogenous samples, such as whole tissues or even whole embryos, during which context-specific effects will be missed through signal averaging across cell populations, particularly in the context of embryogenesis. Isolation of relevant cell-types by FACS can in some instances circumvent this limitation, but often there are no suitable cell markers available for sorting and, because the population of interest needs to be defined a priori, uncharacterized cell-types or transient cell states will be missed.

As demonstrated in (Cusanovich et al., 2018a) and in this thesis (Chapter 4), sci-ATAC-seq can be used to identify embryonic cell populations from whole-embryo samples without requiring cell sorting. Therefore, it could in principle reveal in which cell-types or cell states a genetic variant has a regulatory impact, while assessing all cells-types of an organism in one experiment. *Drosophila melanogaster* is a great model to test case this principle. A large collection of over 200 fully-sequenced inbred lines generated from wild isolates is available to the community (*Drosophila melanogaster* Genetic Reference Panel (DGRP)) (MacKay et al., 2012). These lines harbor a large frequency of SNPs and structural variants (Huang et al., 2014), and this high degree of genetic diversity makes them particularly suitable for allele-specific studies.

In this chapter, I describe the proof-of-principle of this approach. This work is the result of a great collaboration with multiple colleagues, in particular with Tobias Heinen, who designed and implemented the analyses strategy. Their individual contributions are indicated below and more detailed information is given in section ‘5.7 Contributions’.

By crossing genetically diverse fly stains and performing sci-ATAC-seq in the heterozygous F1 progeny, we demonstrate that single-cell chromatin accessibility profiling can discover allelic imbalances and the cellular context in which genetic variants affect enhancer usage.

5.1 Study design

The project I describe in this chapter is based on the use of F1 hybrid embryos, which are obtained by crossing inbred, genetically diverse parent lines (Figure 35). Over 200 fully-sequenced fly lines are available to the community from the DGRP collection (MacKay et al., 2012). The great advantage of this approach is that it enables to disentangle effects due to sequence variation (in *cis*) from effects due to the cellular environment (in *trans*). In the F1 embryos, the maternal and paternal alleles reside in the same cellular environment. Therefore, *trans* effects can be excluded and phenotypic differences can be assumed to be due to *cis* variation in the alleles sequence.

To maximize the recovery of heterogenous effects on enhancer activity, I profiled single-cell chromatin accessibility at three key stages of *Drosophila melanogaster* embryogenesis, 2-4, 6-8 and 10-12 hours (Figure 35). These time points roughly correspond to the stages of multipotency, cell fate specification and terminal differentiation of the embryonic tissues, respectively. Therefore, this time course captures regulatory variation occurring during key developmental transitions and the resulting dataset should encompass enough diverse cell-types and cell states to enable the identification of effects that are context-specific (Figure 35).

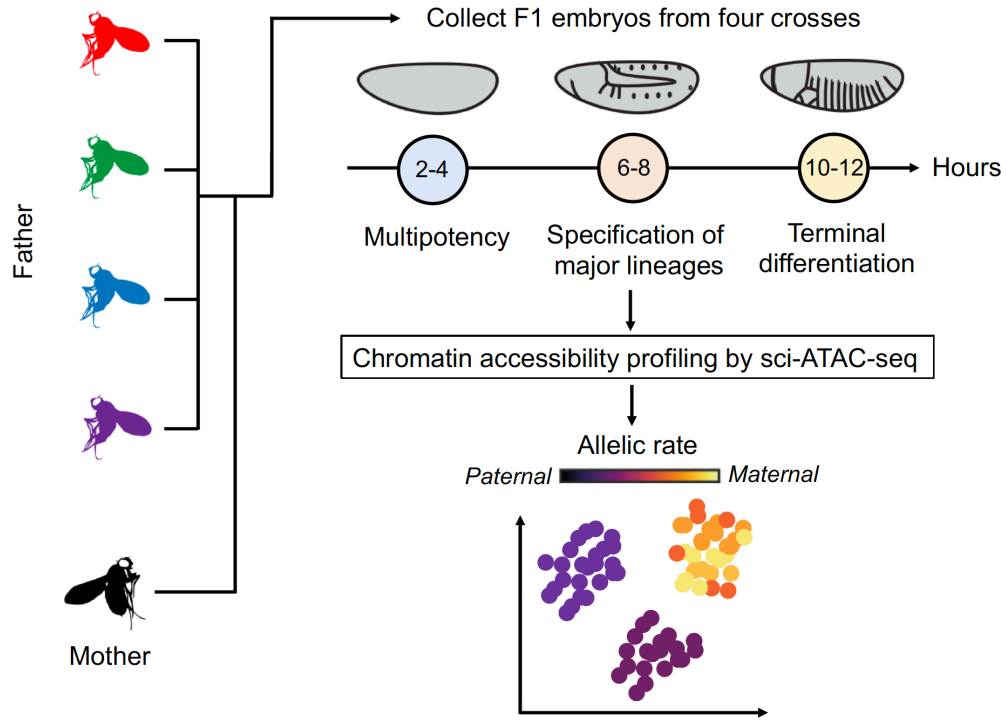

Figure 35. Experimental schematic.

Single-cell chromatin accessibility profiling was performed at three developmental time points of *Drosophila melanogaster* embryogenesis, for F1 hybrid embryos laid by a common maternal line after mating with four genetically distinct paternal lines. For each cell, allelic bias in chromatin accessibility can be quantified and the cell-type and cell state dependency assessed.

Adapted from (Floc'hlay et al., 2021) under Attribution 4.0 International (CC BY 4.0).

5.2 Generation of a high quality sci-ATAC-seq dataset for F1 hybrid embryos

The computational analyses described in this section were performed by Tobias Heinen, unless otherwise stated.

To profile single-cell chromatin accessibility over the time course described above, I extracted nuclei from four collections of F1 hybrid embryos at the indicated time points. The F1 embryos were kindly donated by Bingqing Zhao, a former PhD student in the Furlong lab, who had previously collected them by mating four diverse paternal strains with one common maternal line (Figure 35), in the context of an independent project (Floc'hlay et al., 2021). I performed sci-ATAC-seq on the nuclei to generate a total of 12 sequencing libraries (3 time points x 4 crosses). Given the large amount of Tn5

enzyme and reagents needed to generate this dataset, I made use of the homemade sci-ATAC-seq protocol I implemented (described in Chapter 3). Because of the scale of the project, the sci-ATAC-seq libraries were generated with the assistance of James Reddington, a former staff scientist in the lab.

Computation methods to process and analyze single-cell ATAC-seq data are available (Chen et al., 2019), and in this thesis I demonstrated the use of some of them to analyze chromatin accessibility in both wild-type and mutant *Drosophila* embryos (Chapter 4). However, the scope of this project requires the integration of allele-specific quantifications in order to estimate allelic imbalances in single cells, and tools for this task are still lacking. For this reason, we initiated a collaboration with Tobias Heinen, a PhD student in the Stegle lab, who performed the sci-ATAC-seq data analysis and designed a new tool to integrate allelic counts (scDALI, described below). Tobias and I met frequently to discuss both biological and computational questions, implementation of the analysis strategy, interpretation of results and new directions for further analysis.

Initial QC of the sequenced libraries (performed by Tobias) indicated that a total of 35,485 cells were recovered (average of 9,000 cells per cross) with a median coverage of ~10,000 unique reads per-cell. The dataset is of high quality, as demonstrated by the fact that the sci-ATAC-seq libraries display the expected nucleosomal pattern (Figure 36a) and the accessibility profiles are highly correlated with a published time-matched sci-ATAC-seq dataset (Figure 36b) (Cusanovich et al., 2018a). Therefore, the homemade sci-ATAC-seq protocol I implemented to generate the libraries leads to high quality data and enables identification of the expected cell populations.

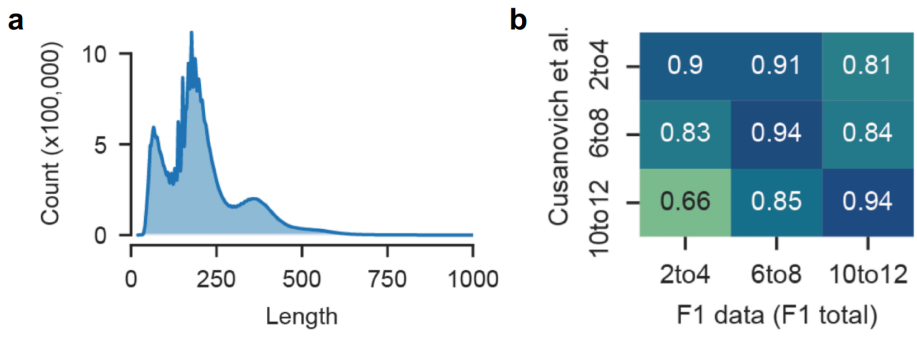

Figure 36. Quality control of sci-ATAC-seq libraries.

(a) Fragment size distribution of the sci-ATAC-seq libraries, displaying the expected nucleosome banding pattern.

(b) Pearson correlation of aggregated chromatin accessibility between the combined F1 datasets and a published time-matched sci-ATAC-seq from a reference *Drosophila melanogaster* line (Cusanovich et al., 2018a).

Tobias Heinen performed the analysis of the sci-ATAC-seq libraries and generated all the plots displayed in this figure.

Dimensionality reduction and clustering on the combined libraries (from all time points and all crosses), revealed the dynamics of the chromatin accessibility landscape (Figure 37). Developmental time is a major driver in the cell clustering and cells arrange gradually according to their developmental progression (Figure 37a). To annotate the cell-types, Tobias identified significantly over-represented tissue-terms within each cluster using two extensive resources: (1) curated enhancers with embryonic activity *in vivo* (Bonn et al., 2012a; Kvon et al., 2014; Rivera et al., 2019) and (2) gene expression throughout embryogenesis (Tomancak et al., 2002). I then used the significant terms identified by Tobias to make a consensus list of cell-type annotations for each cluster. This process identified the major embryonic cell populations, including the ectoderm, nervous and muscle system (Figure 37b). The identified cellular populations are consistent with the ones previously found in a time-matched *Drosophila melanogaster* embryonic dataset (Cusanovich et al., 2018a).

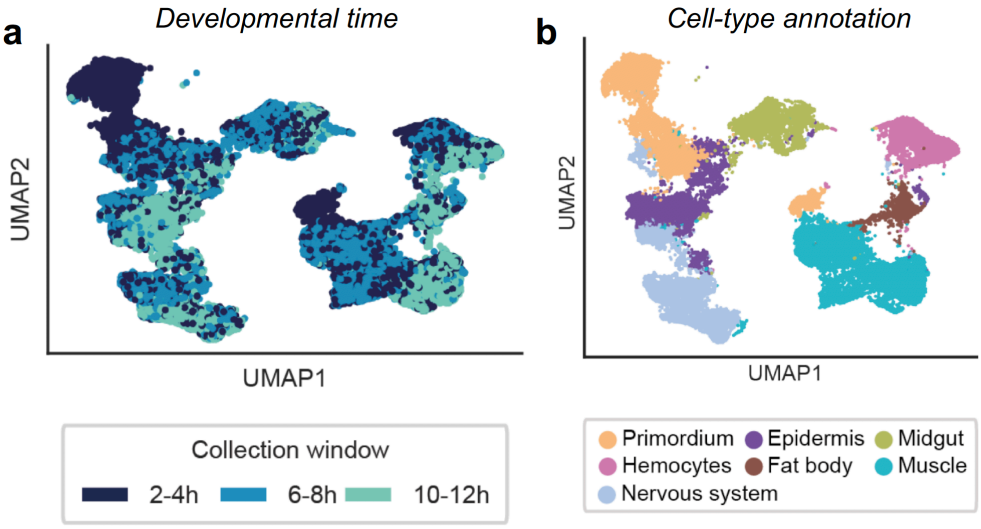

Figure 37. Visualization of the chromatin accessibility landscape and cell-type annotation.

(a) UMAP visualization of all the cells profiled from the 12 sci-ATAC-seq libraries (3 time points by 4 F1 crosses). Each dot represents a cell, colored by the time point of the embryo collections. Cells are clustered based on the similarity in their chromatin accessibility.

(b) Same as (a), cells are colored by the inferred cell-type annotation.

Tobias Heinen performed the analysis of the sci-ATAC-seq libraries and generated all the plots displayed in this figure.

5.3 Context-specific allelic imbalance is common in *Drosophila* embryogenesis

The computational analyses described in this section were performed by Tobias Heinen, unless otherwise stated.

There are currently no published tools that can handle allele-specific scATAC data. To address this need, Tobias Heinen developed scDALI (**s**ingle-**c**ell **d**ifferential **al**lelic **i**mbalance), a tool for identifying and characterizing allelic imbalance from single-cell measurements. Briefly, scDALI integrates allele-specific counts and uses a Bayesian statistical framework to model allelic imbalances at the level of single cells. Importantly, scDALI can distinguish between homogeneous (pervasive across cells) and heterogeneous (cell-type or cell state specific) allelic imbalances. With this approach, it is possible to pin-point context-specific genetic variation to individual cell-types and states, thus revealing genetic effects that impact one tissue or cell-type and not another (Figure 35).

Application of scDALI to test for allelic imbalance across all crosses and time points identified 7,823 regions that display allelic imbalance of any type, of which 415 are heterogenous and variable across cell-types. This analysis shows that there are there hundreds of cell-type specific genetic effects and that heterogenous imbalance is a common occurrence during *Drosophila* embryogenesis. This finding highlights the importance of implementing strategies based on single-cell measurements, rather than on bulk profiling.

5.4 Characterization of regions displaying heterogenous allelic imbalance

The computational analyses described in this section were performed by Tobias Heinen, unless otherwise stated. I manually inspected all regions with predicted allelic imbalance and selected the examples presented in this section.

To better understand the biological context of the imbalanced regions identified by scDALI, I annotated them by searching for overlaps with a collection of curated embryonic enhancers (Bonn et al., 2012a; Kvon et al., 2014; Rivera et al., 2019) and with a large compendium of molecular features profiled by the Furlong lab (TF binding sites, DNase peaks, predicted CRMs). In particular, I focused on characterizing and manually inspecting the 415 regions displaying heterogenous allelic imbalance.

I found many examples of lineage-specific enhancers being affected by strong allelic imbalance. For example, region chr3R:22877489-22878489 encompasses a characterized enhancer that is active in the embryonic nervous system. This region is predominantly accessible in the nervous system and the accessibility is strongly biased for the paternal allele (Figure 38a-c). On the contrary, cells from other lineages do not display a detectable allelic imbalance (Figure 38b,c).

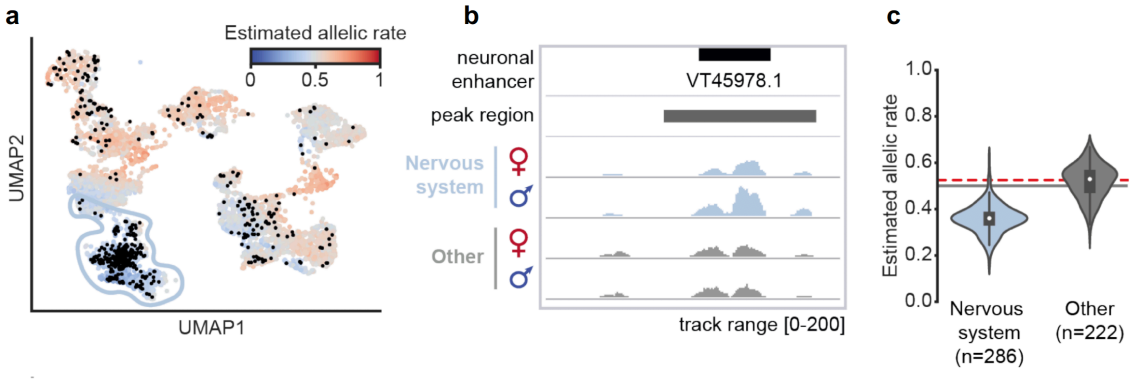

Figure 38. Example of allelic imbalance affecting a lineage-specific enhancer.

(a) UMAP visualization of all the cells profiled from the 12 sci-ATAC-seq libraries (3 time points by 4 F1 crosses). Each dot represents a cell, colored by the allelic rate (maternal allele accessibility over total accessibility) measured at region chr3R:22877489-22878489. Cells with an estimated allelic rate > 0.5 (red) are biased for the maternal allele, cells < 0.5 (blue) are biased for the paternal. Black dots indicate cells with enough counts to permit the allelic rate estimation. The nervous system cluster (as identified in Figure 37b) is circled in light blue.
(b) Genome browser tracks displaying the aggregated chromatin accessibility profiles around region chr3R:22877489-22878489 for the nervous system and the other embryonic populations, split by maternal and paternal allele.
(c) Violin plots showing the distribution of estimated allelic rates per cell, for the nervous system and the other embryonic populations. The number of cells for which the allelic rate was measured is indicated in parenthesis. White dot in the boxplots indicates the median allelic rate. The grey line indicates 0.5 allelic rate, the red line indicates the overall allelic rate estimated across all cells.
Tobias Heinen analyzed the sci-ATAC-seq data, implemented scDALI and generated all the plots displayed in this figure.

Moreover, I also found cases of enhancers displaying opposing allelic imbalance in different tissues. For example, region chr2R:13675707-13676707, was previously identified as a muscle and nervous system putative enhancer based on tissue-specific accessibility data (Reddington et al., 2020) and it is reported by scDALI as a site of strong heterogenous imbalance. Inspection of the region reveals that accessibility is biased for the maternal allele in the muscle and, opposingly, for the paternal allele in the nervous system (Figure 39a-c). The pattern of allelic imbalance is additionally more

complex, as this region is also maternally biased in the primordium of the early embryo (Figure 39a, c).

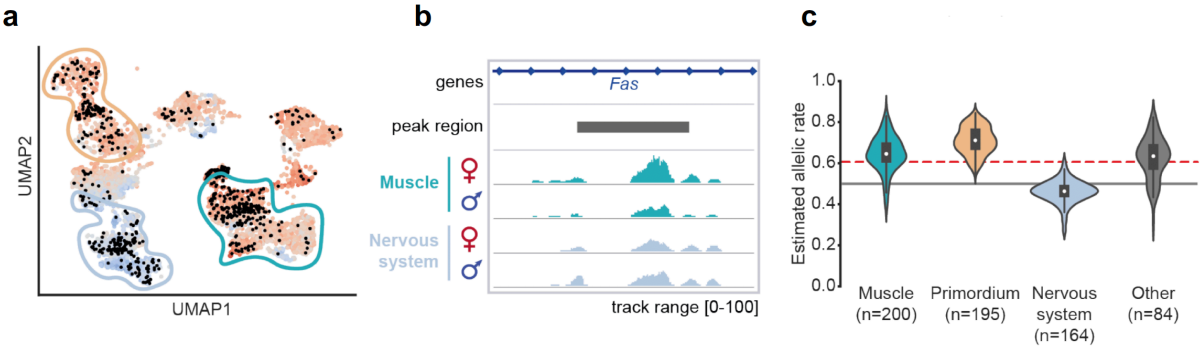

Figure 39. Example of opposing allelic imbalance in different lineages.

(a) UMAP visualization of all the cells profiled from the 12 sci-ATAC-seq libraries (3 time points by 4 F1 crosses). Each dot represents a cell, colored by the allelic rate (maternal allele accessibility over total accessibility) measured at region chr2R:13675707-13676707. Cells with an estimated allelic rate > 0.5 (red) are biased for the maternal allele, cells < 0.5 (blue) are biased for the paternal. Black dots indicate cells with enough counts to permit the allelic rate estimation. The nervous system, the muscle system and the primordium (as identified in Figure 37b) are circled in light blue, green and orange, respectively.

(b) Genome browser tracks displaying the aggregated chromatin accessibility profiles around region chr2R:13675707-13676707 for the muscle and the nervous system, split by maternal and paternal allele.

(c) Violin plots showing the distribution of estimated allelic rates per cell, for the muscle and the muscle, primordium, nervous system and the other embryonic populations. The number of cells for which the allelic rate was measured is indicated in parenthesis. White dot in the boxplots indicates the median allelic rate. The grey line indicates 0.5 allelic rate, the red line indicates the overall allelic rate estimated across all cells.

Tobias Heinen analyzed the sci-ATAC-seq data, implemented scDALI and generated all the plots displayed in this figure.

5.5 On-going experimental validation of heterogeneously imbalanced regions

The spatiotemporal activity of enhancers during embryogenesis can be determined by performing *in vivo* reporter assays. A reporter construct is built by cloning the test enhancer upstream of a reporter gene (*GFP*, *lacZ*) with a minimal promoter, and the construct is then integrated into the genome to generate stable enhancer-reporter fly lines. If the enhancer is capable of activating transcription, expression of the reporter gene is detected (for example by RNA fluorescence in-situ hybridization) in embryos. This approach has already been successfully applied to test spatiotemporal enhancer activity predicted from sci-ATAC-seq data (Cusanovich et al., 2018a).

I decided to generate enhancer-reporter lines for five cases with predicted allelic imbalance by scDALI, with the goal to test if the imbalance leads to changes in the enhancer's activity in the predicted cell-type/state. For each test case, I cloned two versions of the reporter construct, one with the paternal and one with the maternal allele. The enhancer-reporter lines with the integrated constructs have been generated and I am currently doing the subsequent fly crosses necessary to establish the homozygous transgenic lines. Once the homozygous stocks are established, I will then amplify the lines, collect embryos and perform *in-situ* hybridization against the reporter gene, to validate the predicted changes in enhancer activity between the two alleles.

The five regions I selected for testing display a mixture of different effects:

- Region Chr3R:22,877,489-22,878,489 (Figure 38) and Chr2L:8,054,992-8,055,992 are a neuronal and a muscle specific enhancer respectively, and both display strong allelic imbalance within the cellular cluster of their activity. I expect to detect enhancer activity in the corresponding tissue with one allele (paternal for Chr3R:22,877,489-22,878,489, maternal for Chr2L:8,054,992-8,055,992) and a reduction or complete loss of enhancer activity with the opposite allele.
- Region Chr2R:13,675,707-13,676,707 is a neuronal and a muscle putative enhancer, identified from tissue-specific DHS data, that displays opposing allelic imbalance in the two tissues (Figure 39). As predicted by scDALI, I expect

the paternal allele to lead to enhancer activity in the nervous system and the maternal allele in the muscle system.

- Region Chr2R:13,675,707-13,676,707 is a putative enhancer that displays intra-lineage allelic imbalance within the muscle population, which seems to be associated with temporal progression (Figure 40a). I expect muscle enhancer activity to be driven by the paternal allele at early embryonic stages and then by the maternal allele at later stages.
- Region Chr3L:7,988,664-7,989,664 is a putative enhancer that displays pervasive homogenous maternal imbalance across all cell-types (Figure 40b). The prediction is that the maternal allele, but not the paternal, drives enhancer activity in most embryonic populations.

If the results of these predictions hold true, this work will showcase the complex different ways by which genetic variation can impact enhancer activity and gene expression - having specific affects in one tissue or time point or opposing effects.

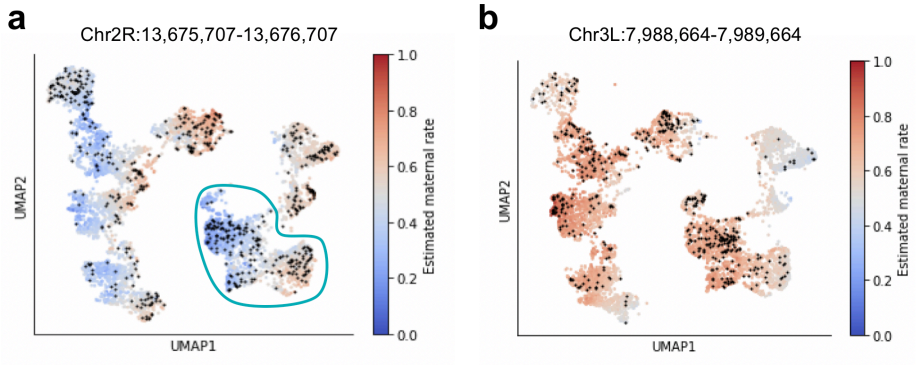

Figure 40. Additional examples of allelic imbalances selected for validation.

(a) UMAP visualization of all the cells profiled from the 12 sci-ATAC-seq libraries (3 time points by 4 F1 crosses). Each dot represents a cell, colored by the allelic rate (maternal allele accessibility over total accessibility) measured at region Chr2R:13,675,707-13,676,707. Cells with an estimated allelic rate > 0.5 (red) are biased for the maternal allele, cells < 0.5 (blue) are biased for the paternal. Black dots indicate cells with enough counts to permit the allelic rate estimation. The muscle system (as identified in Figure 37b) is circled in green.

(b) Same as (a) but for region Chr3L:7,988,664-7,989,664. Chromatin accessibility at this region is homogenously biased for the maternal allele.
Tobias Heinen analyzed the sci-ATAC-seq data, implemented scDALI and generated all the plots displayed in this figure.

5.6 Discussion and Conclusions

Enhancers are core components of developmental gene regulatory networks, yet their necessity and functionality are remarkably difficult to assess, in part because of the practical limitations deleting them or mutating them at large-scale in multicellular organisms. Natural sequence variation offers a means to study the effect of large-scale perturbations of enhancers during embryogenesis. As enhancer activity is cell-type specific, the effect of genetic variants within them will also vary across cell-types and cellular states. However, capturing this heterogenous genetic variation is challenging as suitable methods are lacking. One approach, which consists in using FACS to isolate purified cell-types from complex samples, is often not possible due to the number of samples required or because of the lack of suitable markers, particularly for developmental transitions. For this reason, the vast majority of studies assessing the functional impact of genetic variation during embryogenesis used whole embryos, including previous studies in the Furlong lab assessing the impact of genetic variation on enhancer activity (Floc'hlay et al., 2021), promoter activity (Schor et al., 2017) and gene expression (Cannavò et al., 2017). Such studies that relied on bulk profiling methods cannot assess the context-specific effect of genetic variants in enhancers.

In this chapter, I show together with my collaborator, Tobias, that single-cell chromatin accessibility profiling not only removes the need to isolate cell-types from heterogenous samples, but it also enables to detect the impact of heterogenous imbalances on chromatin accessibility. Profiling over 35,000 single-cells in F1 genetic hybrid embryos at three stages of *Drosophila* embryogenesis was sufficient to reconstruct the developmental chromatin accessibility landscape and robustly identify all the major embryonic populations without the use of FACS purification. Application of scDALI, a tool designed to exploit allele-specific accessibility counts to infer both homogenous and heterogenous allelic imbalances at the single-cell level, detects

thousands of imbalanced regions, hundreds of which are heterogenous and vary by cell-type. Thus, heterogenous imbalance is a widespread occurrence during *Drosophila* embryogenesis. Inspection of the heterogeneously imbalanced regions revealed many instances in which enhancers are affected, even cases in which enhancers display opposing imbalance in different tissues. Many of these context-specific effects were missed in the lab's previous study applying bulk profiling (Floc'hlay et al., 2021), highlighting the importance to implement strategies based on single-cell measurements.

The approach presented here is not limited to *Drosophila* and should easily be extendable to other model organisms. It is applicable to both F1 crosses of genetically diverse inbred lines or to large sample collections from individuals in a population. Moreover, the approach is not limited to chromatin accessibility, but should be applicable to any single-cell sequencing method that can identify cell-types or cellular states and capture allelic-specific measurements. For example, the impact of genetic variations on chromatin state or TF binding could be studied with single-cell ChIP-seq and the impact on gene expression by single-cell RNA-seq, all of which can be flexibly integrated by scDALI.

Going forward, the allelic imbalance characterized here should be further examined to uncover how they mechanistically impact chromatin accessibility and enhancer function. Scanning for alterations in the TF motifs present in the imbalanced regions could reveal an impact on TF binding, which could then be validated using both *in vitro* and *in vivo* assays. Profiling additional modalities, such as the chromatin state by scChIP-seq and gene expression by scRNA-seq, would allow to more extensively characterize the impact of genetic variation on other molecular layers beyond chromatin accessibility.

5.7 Contributions

I performed the experimental work presented in this chapter, unless otherwise stated. I participated in regular meetings with Tobias Heinen, Dr. Oliver Stegle and Dr. Eileen Furlong to discuss the project conceptualization, the experimental and the computational strategy and the interpretation of the results.

Tobias Heinen (*PhD student, Stegle lab*) analyzed the sci-ATAC-seq data and designed and implemented scDALI. I annotated the list of imbalanced regions generated by scDALI using data resources generated by our lab (TF ChIP, DNase, enhancer catalogues, etc.) and identified candidates for experimental validation.

Dr. James Reddington (*former staff scientist, Furlong lab*) provided assistance with the implementation of sci-ATAC-seq and generation of the libraries.

Dr. Bingqing Zhao (*former PhD student, Furlong lab*) collected the F1 hybrid embryos.

Staff at the EMBL Protein Expression and Purification Facility produced the Tn5 enzyme used for sci-ATAC-seq.

Staff at the EMBL Genecore Facility performed all sequencing. I prepared and submitted all the libraries to the facility for sequencing.

Dr. Eileen Furlong (*Group Leader, Furlong lab*) and Dr. Oliver Stegle (*Group Leader, Stegle lab*) supervised the project, made intellectual contributions and provided funding.

6 Final Remarks and Future Perspectives

The characterization of the genetic mechanisms that control developmental processes remains a central goal in biology. During my thesis, I tried to gain insight into those underlying regulatory mechanisms. For that, I developed three interconnected projects that take advantage of single-cell chromatin accessibility profiling to characterize the regulatory landscape of developing *Drosophila* embryos and asses the alterations caused by perturbation of the regulatory networks.

In the first project (Chapter 3), I optimized a protocol to profile single-cell chromatin accessibility by sci-ATAC-seq in fixed *Drosophila* embryos. With improvements in the reaction conditions, the protocol reproducibly yields high quality libraries with a per-cell coverage that is comparable to that obtained in other model systems. I also implemented a version of the protocol that uses homemade Tn5 enzyme and cheaper PCR reagents, making the protocol more cost-effective and suitable for large-scale projects.

In the second project (Chapter 4), I harnessed the optimized sci-ATAC-seq protocol to profile chromatin accessibility for over 20,000 cells at high temporal resolution during *Drosophila* mesoderm development in wild-type embryos. This dataset captures the establishment of regulatory landscapes in nascent embryonic mesoderm and its progression along developmental time. Through the integration of extensive information on characterized enhancers' activity and transcription factor occupancy, I could resolve diverse muscle sub-populations and rare cell types, including the heart. I demonstrate that this information can reconstruct multiple developmental trajectories corresponding to lineage choices along different muscle primordia, and uncover the underlying regulatory programs, including the transcription factors involved. To study the functional impact of such transcription factors on enhancer regulation and cell fate decisions, together with my collaborators, I set up a streamlined approach to systematically profile chromatin accessibility and *de-novo* genotype mutant cells coming from embryo collections of mixed genotypes. sci-ATAC-seq profiling of mutant embryos for four essential TFs for muscle development revealed strikingly different outcomes in cell fate. While one mutant (*Mef2*) results in abnormal 'new' cell states,

other TF mutants appear to block cell populations much earlier along their developmental trajectory. Further characterization of the mutants uncovered new subtle phenotypes, besides previously described ones, and also provided information of the TF role at enhancers. This approach can thereby reveal the functional impact of developmental mutants at both a cellular and molecular level at a scale and resolution that was not feasible before.

In the third project (Chapter 5), I once more perturbed *Drosophila* developmental regulatory networks assessed the outcome with sci-ATAC-seq. Because large-scale mutagenesis targeting regulatory elements is poorly scalable beyond cell culture, I exploited natural sequence variation as a means for large-scale enhancer perturbation in *cis*. Together with my collaborator Tobias Heinen, I show that, by profiling chromatin accessibility in hybrid embryos that were obtained from mating genetically diverse parent lines, genetic variants can be mapped to the cellular context in which they have an impact. I find many instances of genetic variants having heterogenous effects across cell-types and affect enhancers, including cases in which enhancers are affected in opposing ways in different tissues. Many of these effects cannot be captured by bulk sequencing methods, highlighting the importance of approaching these questions at single-cell resolution.

A longstanding challenge in the field is to move beyond the discovery of new regulatory regions and to accurately infer their function, how they relate to other molecular layers and how they influence gene expression. To begin addressing these questions, in my thesis, I developed two approaches that enable to assess the impact of perturbations of regulatory networks using sci-ATAC-seq. Chromatin accessibility profiling is a great tool to identify the repertoire of regulatory elements available to the cell and, as demonstrated in my thesis, it enables to track changes in enhancers usage and identify cell-types in complex heterogenous samples. However, by itself chromatin accessibility fails to reveal how regulatory regions function - whether they act as a promoter, enhancer or other regulatory element, and their current activity state. While it has been shown that prediction of enhancer activity from chromatin accessibility alone can be remarkably high, particularly for regions identified with single-cell profiling (Bravo González-Blas et al., 2020; Cusanovich et al., 2018a), accessibility does not always translate into activity. For example, the binding of a transcriptional repressor

would also result in locally increased chromatin accessibly but not in enhancer activation. Because ATAC-seq does not directly provide the identity of the factors occupying the accessible regions, but they need to be inferred from the underlying sequence, the prediction power is reduced.

Overcoming this limitation requires the integration of chromatin accessibility with more single-cell profiling methods that interrogate additional molecular layers. For example, the timing of enhancer activation could be more precisely predicted by combining chromatin accessibility with the profiling of H3K27ac, polymerase II and TFs occupancy (Bonn et al., 2012a; Zinzen et al., 2009), using methods such as single-cell Cut&Tag (Kaya-Okur et al., 2019) and CoBATCH (Wang et al., 2019). While these molecular layers are usually measured separately, new methods that enable multi-omic measurements in the same cell are rapidly emerging, with protocols already existing for the joint profiling of chromatin accessibility and either the transcriptome (Cao et al., 2018; Ma et al., 2020), protein-abundance (Chen et al., 2018) or DNA methylation (Argelaguet et al., 2019; Pott, 2016). For example, having both chromatin accessibility and gene expression from the same cell, should facilitate the direct association of enhancer to their target genes. These methods, therefore, hold the promise to greatly improve prediction power.

An important way to uncover the function of different components of regulatory networks is to perturb their expression. In chapter 4 of my thesis, I applied the classic genetic approach of generating TF loss-of-function mutants and assessed their role on cell fate and enhancer regulation. For example, when I mutated *biniou*, a TF required for visceral muscle specification, I could observe a complete loss of the visceral muscle lineage (for details please refer to section 4.10: 'Loss of *tinman*, *bagpipe* and *biniou* differentially alters cellular composition'). A small population of progenitor cells that appeared normal in their chromatin accessibility was still present, but was unable to continue developmental progression because the specifying TF was missing. Because of this I could not characterize the chromatin alterations induced by the TF removal and I could not assess the function of the TF in the later differentiation steps.

This approach would greatly benefit from integration with a system for tunable depletion of the TF. Optogenetics systems enable to deplete a TF from the nucleus

and regulate its concentration, by fusing the factor with a light inducible nuclear localization tag (de Mena et al., 2018; Niopek et al., 2014). Integration of this system would enable to remove Biniou from the nucleus after the visceral lineage is specified, and thus enable the investigation of its role into the later stages of visceral differentiation. Moreover, the impact of a whole range of TF depletions with increasing intensity could be characterized. Starting from the unaffected progenitors, this process should enable to reconstruct the trajectory taken by the cells as they progress towards their altered phenotype. In the analysis of the *Mef2* mutant (for details please refer to section 4.11: 'Removal of Mef2 affects chromatin accessibility at hundreds of regulatory regions') I identified 197 genomic regions whose chromatin accessibility is dependent on the binding of Mef2, however I could not discern the temporal order by which these regions require the factor. This could be revealed by performing a time course of Mef2 depletion from the nucleus along lineage development, to dissect at high resolution the temporal progression by which the regions are bound by the TF.

A major goal driving the reconstruction of developmental regulatory landscapes is to build predictive models of gene regulation. The datasets and the approaches presented in my thesis can contribute to this in at least three ways. First, the datasets I generated provide high-quality single-cell chromatin accessibility profiles that can be incorporated into models exploring *Drosophila* embryogenesis. Second, the high resolution of my datasets can be used to refine previously generated bulk data, for example TF occupancy data, to a more precise temporal window and cellular context, and thus contribute to improving existing models. Third, the approaches described here enable to assess how perturbations of both TFs and enhancers differentially affect cell populations in complex samples at a fine-grained resolution, which is crucial for the creation of accurate predictive models. Ultimately, these and similar approaches, in combination with the growing capability of single-cell technologies, will make it possible to build virtual embryo models that can predict cell-resolved spatiotemporal gene expression and precisely simulate the impact of perturbations in the regulatory networks.

7 References

Abmayr, S.M., and Keller, C.A. (1997). Drosophila Myogenesis and insights into the Role of nautilus. Curr. Top. Dev. Biol. *38*, 35–80.

Adan, A., Alizada, G., Kiraz, Y., Baran, Y., and Nalbant, A. (2017). Flow cytometry: basic principles and applications. Crit. Rev. Biotechnol. *37*, 163–176.

Alberts, B., Johnson, A., and Lewis, J. (2002). Molecular Biology of the Cell. Drosophila and the Molecular Genetics of Pattern Formation: Genesis of the Body Plan. (New York: Garland Science).

Argelaguet, R., Clark, S.J., Mohammed, H., Stapel, L.C., Krueger, C., Kapourani, C.A., Imaz-Rosshandler, I., Lohoff, T., Xiang, Y., Hanna, C.W., et al. (2019). Multi-omics profiling of mouse gastrulation at single-cell resolution. Nature *576*, 487–491.

Azpiazu, N., and Frasch, M. (1993). Tinman and bagpipe: Two homeo box genes that determine cell fates in the dorsal mesoderm of Drosophila. Genes Dev. *7*, 1325–1340.

Azpiazu, N., Lawrence, P.A., Vincent, J.P., and Frasch, M. (1996). Segmentation and specification of the Drosophila mesoderm. Genes Dev. *10*, 3183–3194.

Bae, Y.K., Macabenta, F., Curtis, H.L., and Stathopoulos, A. (2017). Comparative analysis of gene expression profiles for several migrating cell types identifies cell migration regulators. Mech. Dev. *148*, 40–55.

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

Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-Resolution Profiling of Histone Methylations in the Human Genome. Cell *129*, 823–837.

Baylies, M.K., and Bate, M. (1996). twist: A myogenic switch in Drosophila. Science (80-.). *272*, 1481–1484.

Behera, V., Evans, P., Face, C.J., Hamagami, N., Sankaranarayanan, L., Keller, C.A., Giardine, B., Tan, K., Hardison, R.C., Shi, J., et al. (2018). Exploiting genetic variation to uncover rules of transcription factor binding and chromatin accessibility. Nat. Commun. *9*.

Blow, M.J., McCulley, D.J., Li, Z., Zhang, T., Akiyama, J.A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., et al. (2010). ChIP-seq identification of weakly conserved heart enhancers. Nat. Genet. *42*, 806–812.

Bodmer, R. (1993). The gene tinman is required for specification of the heart and visceral muscles in Drosophila. Development *118*, 719–729.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

Bonn, S., and Furlong, E.E.M. (2008). cis-Regulatory networks during development: a view of Drosophila. Curr. Opin. Genet. Dev. *18*, 513–520.

Bonn, S., Zinzen, R.P., Girardot, C., Gustafson, E.H., Perez-Gonzalez, A., Delhomme, N., Ghavi-Helm, Y., Wilczyński, B., Riddell, A., and Furlong, E.E.M. (2012a). Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. Nat. Genet. *44*, 148–156.

Bonn, S., Zinzen, R.P., Perez-Gonzalez, A., Riddell, A., Gavin, A.C., and Furlong, E.E.M. (2012b). Cell type-specific chromatin immunoprecipitation from multicellular complex samples using bits-chip. Nat. Protoc. *7*, 978–994.

Bour, B.A., O'Brien, M.A., Lockwood, W.L., Goldstein, E.S., Bodmer, R., Taghert, P.H., Abmayr, S.M., and Nguyen, H.T. (1995). Drosophila MEF2, a transcription factor that is essential for myogenesis. Genes Dev. *9*, 730–741.

Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z., Furey, T.S., and Crawford, G.E. (2008). High-Resolution Mapping and Characterization of Open Chromatin across the Genome. Cell *132*, 311–322.

Bozek, M., Cortini, R., Storti, A.E., Unnerstall, U., Gaul, U., and Gompel, N. (2019). ATAC-seq reveals regional differences in enhancer accessibility during the establishment of spatial coordinates in the Drosophila blastoderm. Genome Res. *29*, 771–783.

Bravo González-Blas, C., Minnoye, L., Papasokrati, D., Aibar, S., Hulselmans, G., Christiaens, V., Davie, K., Wouters, J., and Aerts, S. (2019). cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. Nat. Methods *16*, 397–400.

Bravo González-Blas, C., Quan, X., Duran-Romaña, R., Taskiran, I.I., Koldere, D., Davie, K., Christiaens, V., Makhzami, S., Hulselmans, G., Waegeneer, M., et al. (2020). Identification of genomic enhancers through spatial integration of single-cell transcriptomics and epigenomics. Mol. Syst. Biol. *16*, 1–32.

Briggs, J.A., Weinreb, C., Wagner, D.E., Megason, S., Peshkin, L., Kirschner, M.W., and Klein, A.M. (2018). The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. Science (80-.). *360*.

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

Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P., Chang, H.Y., and Greenleaf, W.J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. Nature *523*, 486–490.

Cannavò, E., Koelling, N., Harnett, D., Garfield, D., Casale, F.P., Ciglar, L., Gustafson, H.E., Viales, R.R., Marco-Ferreres, R., Degner, J.F., et al. (2017). Genetic variants regulating expression levels and isoform diversity during embryogenesis. Nature *541*, 402–406.

Cao, J., Cusanovich, D.A., Ramani, V., Aghamirzaie, D., Pliner, H.A., Hill, A.J., Daza, R.M., McFaline-Figueroa, J.L., Packer, J.S., Christiansen, L., et al. (2018). Joint profiling of chromatin accessibility and gene expression in thousands of single cells. Science (80-.). *1385*, 1380–1385.

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

Chen, H., Lareau, C., Andreani, T., Vinyard, M.E., Garcia, S.P., Clement, K., Andrade-Navarro, M.A., Buenrostro, J.D., and Pinello, L. (2019). Assessment of computational methods for the analysis of single-cell ATAC-seq data. Genome Biol. *20*, 1–25.

Chen, X., Litzenburger, U.M., Wei, Y., Schep, A.N., LaGory, E.L., Choudhry, H., Giaccia, A.J., Greenleaf, W.J., and Chang, H.Y. (2018). Joint single-cell DNA accessibility and protein epitope profiling reveals environmental regulation of epigenomic heterogeneity. Nat. Commun. *9*.

Chung, C.Y., Ma, Z., Dravis, C., Preissl, S., Poirion, O., Luna, G., Hou, X., Giraddi, R.R., Ren, B., and Wahl, G.M. (2019). Single-Cell Chromatin Analysis of Mammary Gland Development Reveals Cell-State Transcriptional Regulators and Lineage Relationships. Cell Rep. *29*, 495–510.

Ciglar, L., and Furlong, E.E. (2009). Conservation and divergence in developmental networks: a view from Drosophila myogenesis. Curr. Opin. Cell Biol. *21*, 754–760.

Corces, M.R., Trevino, A.E., Hamilton, E.G., Greenside, P.G., Sinnott-Armstrong, N.A., Vesuna, S., Satpathy, A.T., Rubin, A.J., Montine, K.S., Wu, B., et al. (2017). An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. Nat. Methods *14*, 959–962.

Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A., et al. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc. Natl. Acad. Sci. U. S. A. *107*, 21931–21936.

Crittenden, J.R., Skoulakis, E.M.C., Goldstein, E.S., and Davis, R.L. (2018). Drosophila mef2 is essential for normal mushroom body and wing development. Biol. Open *7*.

Cunha, P.M.F., Sandmann, T., Hilary Gustafson, E., Ciglar, L., Eichenlaub, M.P., and Furlong, E.E.M. (2010). Combinatorial binding leads to diverse regulatory responses: Lmd is a tissue-specific modulator of Mef2 activity. PLoS Genet. *6*, 1–11.

Cusanovich, D. a, Daza, R., Adey, A., Pliner, H. a, Christiansen, L., Gunderson, K.L., Steemers, F.J., Trapnell, C., and Shendure, J. (2015). Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. Science *348*, 910–914.

Cusanovich, D.A., Reddington, J.P., Garfield, D.A., Daza, R.M., Aghamirzaie, D., Marco-Ferreres, R., Pliner, H.A., Christiansen, L., Qiu, X., Steemers, F.J., et al. (2018a). The cis-regulatory dynamics of embryonic development at single-cell resolution. Nature *555*, 538–542.

Cusanovich, D.A., Hill, A.J., Aghamirzaie, D., Daza, R.M., Pliner, H.A., Berletch, J.B., Filippova, G.N., Huang, X., Christiansen, L., DeWitt, W.S., et al. (2018b). A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. Cell 1–16.

Daugherty, A.C., Yeo, R.W., Buenrostro, J.D., Greenleaf, W.J., Kundaje, A., and Brunet, A. (2017). Chromatin accessibility dynamics reveal novel functional enhancers in C. elegans. Genome Res. *27*, 2096–2107.

Davidson, E.H., and Douglas, E.H. (2006). Gene regulatory networks and the evolution of animal body plans. Science (80-.). *311*, 796–800.

Davie, K., Janssens, J., Koldere, D., De Waegeneer, M., Pech, U., Kreft, Ł., Aibar, S., Makhzami, S., Christiaens, V., Bravo González-Blas, C., et al. (2018). A Single-Cell Transcriptome Atlas of the Aging Drosophila Brain. Cell *174*, 982–998.

Degner, J.F., Pai, A.A., Pique-Regi, R., Veyrieras, J.B., Gaffney, D.J., Pickrell, J.K., De Leon, S., Michelini, K., Lewellen, N., Crawford, G.E., et al. (2012). DNase-I sensitivity QTLs are a major determinant of human expression variation. Nature *482*, 390–394.

Deng, S., Azevedo, M., and Baylies, M. (2017). Acting on identity: Myoblast fusion and the formation of the syncytial muscle fiber. Semin. Cell Dev. Biol. *72*, 45–55.

Dhall, A., Wei, S., Fierz, B., Woodcock, C.L., Lee, T.H., and Chatterjee, C. (2014). Sumoylated human histone H4 prevents chromatin compaction by inhibiting long-range internucleosomal interactions. J. Biol. Chem. *289*, 33827–33837.

Domcke, S., Hill, A.J., Daza, R.M., Cao, J., O'Day, D.R., Pliner, H.A., Aldinger, K.A., Pokholok, D., Zhang, F., Milbank, J.H., et al. (2020). A human cell atlas of fetal chromatin accessibility. Science (80-.). *370*.

Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C. a, Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Egelhofer, T. a, Minoda, A., Klugman, S., Lee, K., Kolasinska-Zwierz, P., Alekseyenko, A. a, Cheung, M.-S., Day, D.S., Gadel, S., Gorchakov, A. a, et al. (2011). An assessment of histone-modification antibody quality. Nat. Struct. Mol. Biol. *18*, 91–93.

Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43–49.

Farley, E.K., Olson, K.M., and Levine, M.S. (2016). Regulatory principles governing tissue specificity of developmental enhancers. Cold Spring Harb. Symp. Quant. Biol. *80*, 27–32.

Farrell, J.A., Wang, Y., Riesenfeld, S.J., Shekhar, K., Regev, A., and Schier, A.F. (2018). Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. Science (80-.). *360*.

Fenley, A.T., Anandakrishnan, R., Kidane, Y.H., and Onufriev, A. V. (2018). Modulation of nucleosomal DNA accessibility via charge-altering post-translational modifications in histone core. Epigenetics and Chromatin *11*, 1–19.

Floc'hlay, S., Wong, E.S., Zhao, B., Viales, R.R., Thomas-Chollier, M., Thieffry, D., Garfield, D.A., and Furlong, E.E.M. (2021). Cis-acting variation is common across regulatory layers but is often buffered during embryonic development. Genome Res. *31*, 211–224.

Furlong, E.E.M., and Levine, M. (2018). Developmental enhancers and chromosome topology. Science (80-.). *361*, 1341–1345.

Galagan, J.E., Minch, K., Peterson, M., Lyubetskaya, A., Azizi, E., Sweet, L., Gomes, A., Rustad, T., Dolganov, G., Glotova, I., et al. (2013). The Mycobacterium tuberculosis regulatory network and hypoxia. Nature *499*, 178–183.

Ghavi-Helm, Y., Jankowski, A., Meiers, S., Viales, R.R., Korbel, J.O., and Furlong, E.E.M. (2019). Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. Nat. Genet. *51*, 1272–1282.

Giresi, P.G., Kim, J., McDaniell, R.M., Iyer, V.R., and Lieb, J.D. (2007). FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Res. *17*, 877–885.

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

Gratz, S.J., Rubinstein, C.D., Harrison, M.M., Wildonger, J., and O'Connor-Giles, K.M. (2015). CRISPR-Cas9 genome editing in Drosophila. Curr. Protoc. Mol. Biol. *2015*, 31.2.1-31.2.20.

Guenther, M.G., Levine, S.S., Boyer, L.A., Jaenisch, R., and Young, R.A. (2007). A Chromatin Landmark and Transcription Initiation at Most Promoters in Human Cells. Cell *130*, 77–88.

Gurdon, J.B., Laskey, R.A., and Reeves, O.R. (1975). The developmental capacity of nuclei transplanted from keratinized skin cells of adult frogs. J. Embryol. Exp. Morphol. *34*, 93–112.

Haines, J.E., and Eisen, M.B. (2018). Patterns of chromatin accessibility along the anterior-posterior axis in the early Drosophila embryo. PLoS Genet. *14*.

Hanlon, C.D., and Andrew, D.J. (2016). Drosophila FoxL1 non-autonomously coordinates organ placement during embryonic development. Dev. Biol. *419*, 273–284.

Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A., et al. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat. Genet. *39*, 311–318.

Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W., et al. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature *459*, 108–112.

Huang, W., Massouras, A., Inoue, Y., Peiffer, J., Ràmia, M., Tarone, A.M., Turlapati, L., Zichner, T., Zhu, D., Lyman, R.F., et al. (2014). Natural variation in genome architecture among 205 Drosophila melanogaster Genetic Reference Panel lines. Genome Res. *24*, 1193–1208.

Huang, Y., McCarthy, D.J., and Stegle, O. (2019). Vireo: Bayesian demultiplexing of pooled single-cell RNA-seq data without genotype reference. Genome Biol. *20*.

Jakobsen, J.S., Braun, M., Astorga, J., Gustafson, E.H., Sandmann, T., Karzynski, M., Carlsson, P., and Furlong, E.E.M. (2007). Temporal ChIP-on-chip reveals Biniou as a universal regulator of the visceral muscle transcriptional network. Genes Dev. *21*, 2448–2460.

Junion, G., Spivakov, M., Girardot, C., Braun, M., Gustafson, E.H., Birney, E., and Furlong, E.E.M. (2012). A transcription factor collective defines cardiac cell fate and reflects lineage history. Cell *148*, 473–486.

Karaiskos, N., Wahle, P., Alles, J., Boltengagen, A., Ayoub, S., Kipar, C., Kocks, C., Rajewsky, N., and Zinzen, R.P. (2017). The Drosophila embryo at single-cell transcriptome resolution. Science (80-.). *358*, 194–199.

Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S.M., Habegger, L., Rozowsky, J., Shi, M., Urban, A.E., et al. (2010). Variation in transcription factor binding among humans. Science (80-.). *328*, 232–235.

Kaya-Okur, H.S., Wu, S.J., Codomo, C.A., Pledger, E.S., Bryson, T.D., Henikoff, J.G., Ahmad, K., and Henikoff, S. (2019). CUT&Tag for efficient epigenomic profiling of small samples and single cells. Nat. Commun. *10*, 1–10.

Kelsey, G., and Stegle, O. (2017). Learning from the past and predicting the future. Science (80-.). *127*, 2017.

Klapper, R., Stute, C., Schomaker, O., Strasser, T., Janning, W., Renkawitz-Pohl, R., and Holz, A. (2002). The formation of syncytia within the visceral musculature of the Drosophila midgut is dependent on duf, sns and mbc. Mech. Dev. *110*, 85–96.

Kornberg, R.D., and Lorch, Y. (1992). Chromatin structure and transcription. Annu. Rev. Cell Biol. *8*, 563–587.

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P. ru, and Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. Nat. Methods *16*, 1289–1296.

Kouzarides, T. (2007). Chromatin Modifications and Their Function. Cell *128*, 693–705.

Kudron, M.M., Victorsen, A., Gevirtzman, L., Hillier, L.W., Fisher, W.W., Vafeados, D., Kirkey, M., Hammonds, A.S., Gersch, J., Ammouri, H., et al. (2018). The modern resource: genome-wide binding profiles for hundreds of Drosophila and Caenorhabditis elegans transcription factors. Genetics *208*, 937–949.

Kvon, E.Z., Kazmar, T., Stampfel, G., Yáñez-Cuna, J.O., Pagani, M., Schernhuber, K., Dickson, B.J., and Stark, A. (2014). Genome-scale functional characterization of Drosophila developmental enhancers in vivo. Nature *512*, 91–95.

Kwasnieski, J.C., Fiore, C., Chaudhari, H.G., and Cohen, B.A. (2014). High-throughput functional testing of ENCODE segmentation predictions. Genome Res. *24*, 1595–1602.

Lake, B.B., Chen, S., Sos, B.C., Fan, J., Kaeser, G.E., Yung, Y.C., Duong, T.E., Gao, D., Chun, J., Kharchenko, P. V., et al. (2018). Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. Nat. Biotechnol. *36*, 70–80.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Lareau, C.A., Duarte, F.M., Chew, J.G., Kartha, V.K., Burkett, Z.D., Kohlway, A.S., Pokholok, D., Aryee, M.J., Steemers, F.J., Lebofsky, R., et al. (2019). Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. Nat. Biotechnol. *37*, 916–924.

Lee, D.M., and Chen, E.H. (2019). Drosophila Myoblast Fusion: Invasion and Resistance for the Ultimate Union. Annu. Rev. Genet. *53*, 67–91.

Lee, C.K., Shibata, Y., Rao, B., Strahl, B.D., and Lieb, J.D. (2004). Evidence for nucleosome depletion at active regulatory regions genome-wide. Nat. Genet. *36*, 900–905.

Lee, H.-H., Zaffran, S., and Frasch, M. (2006). Develpment of the larval visceral musculature. In Muscle development in Drosophila (ed. H. Sink) (Landes Bioscience, Austin, TX).

Levine, M., and Davidson, E. (2005). Gene regulatory networks for development. Pnas *102*, 4936–4942.

Li, G., and Reinberg, D. (2011). Chromatin higher-order structures and gene regulation. Curr. Opin. Genet. Dev. *21*, 175–186.

Liber, D., Domaschenz, R., Holmqvist, P.H., Mazzarella, L., Georgiou, A., Leleu, M., Fisher, A.G., Labosky, P.A., and Dillon, N. (2010). Epigenetic priming of a Pre-B Cell-Specific enhancer through binding of Sox2 and Foxd3 at the ESC stage. Cell Stem Cell *7*, 114–126.

Lilly, B., Zhao, B., Ranganayakulu, G., Paterson, B.M., Schulz, R.A., and Olson, E.N. (1995). Requirement of MADS domain transcription factor D-MEF2 for muscle formation in Drosophila. Science (80-.). *267*, 688–693.

Lin, S.C., Lin, M.H., Horváth, P., Reddy, K.L., and Storti, R. V. (1997). PDP1, a novel Drosophila PAR domain bZIP transcription factor expressed in developing mesoderm, endoderm and ectoderm, is a transcriptional regulator of somatic muscle genes. Development *124*, 4685–4696.

Liu, Y.H., Jakobsen, J.S., Valentin, G., Amarantos, I., Gilmour, D.T., and Furlong, E.E.M. (2009). A Systematic Analysis of Tinman Function Reveals Eya and JAK-STAT Signaling as Essential Regulators of Muscle Development. Dev. Cell *16*, 280–291.

Ma, S., Zhang, B., Lafave, L.M., Hsu, Y., Regev, A., and Buenrostro, J.D. (2020). Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. Cell *183*, 1103–1116.

MacKay, T.F.C., Richards, S., Stone, E.A., Barbadilla, A., Ayroles, J.F., Zhu, D., Casillas, S., Han, Y., Magwire, M.M., Cridland, J.M., et al. (2012). The Drosophila melanogaster Genetic Reference Panel. Nature *482*, 173–178.

Manolio, T.A. (2010). Genomewide Association Studies and Assessment of the Risk of Disease. N. Engl. J. Med. *363*, 166–176.

Marand, A.P., Chen, Z., Gallavotti, A., and Schmitz, R.J. (2021). A cis-regulatory atlas in maize at single-cell resolution. Cell *184*, 3041-3055.e21.

Margueron, R., Trojer, P., and Reinberg, D. (2005). The key to development: Interpreting the histone code? Curr. Opin. Genet. Dev. *15*, 163–176.

Mbodj, A., Gustafson, E.H., Ciglar, L., Junion, G., Gonzalez, A., Girardot, C., Perrin, L., Furlong, E.E.M., and Thieffry, D. (2016). Qualitative Dynamical Modelling Can Formally Explain Mesoderm Specification and Predict Novel Developmental Phenotypes. PLoS Comput. Biol. *12*, 1–17.

McKay, D.J., and Lieb, J.D. (2013). A Common Set of DNA Regulatory Elements Shapes Drosophila Appendages. Dev. Cell *27*, 306–318.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Garbriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 1297–1303.

de Mena, L., Rizk, P., and Rincon-Limas, D.E. (2018). Bringing Light to Transcription: The Optogenetics Repertoire. Front. Genet. *9*, 1–12.

Mercer, E.M., Lin, Y.C., Benner, C., Jhunjhunwala, S., Dutkowski, J., Flores, M., Sigvardsson, M., Ideker, T., Glass, C.K., and Murre, C. (2011). Multilineage Priming of Enhancer Repertoires Precedes Commitment to the B and Myeloid Cell Lineages in Hematopoietic Progenitors. Immunity *35*, 413–425.

Mikhaylichenko, O., Bondarenko, V., Harnett, D., Schor, I.E., Males, M., Viales, R.R., and Furlong, E.E.M. (2018). The degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription. Genes Dev. *32*, 42–57.

Miller, D.E., Cook, K.R., Arvanitakis, A. V., and Hawley, R.S. (2016). Third chromosome balancer inversions disrupt protein-coding genes and influence distal recombination events in Drosophila melanogaster. G3 Genes, Genomes, Genet. *6*, 1959–1967.

Miller, D.E., Cook, K.R., Hemenway, E.A., Fang, V., Miller, A.L., Hales, K.G., and Hawley, R.S. (2018). The molecular and genetic characterization of second chromosome balancers in Drosophila melanogaster. G3 Genes, Genomes, Genet. *8*, 1161–1171.

Minnoye, L., Marinov, G.K., Krausgruber, T., Pan, L., Marand, A.P., Secchia, S., Greenleaf, W.J., Furlong, E.E.M., Zhao, K., Schmitz, R.J., et al. (2021). Chromatin accessibility profiling methods. Nat. Rev. Methods Prim. *1*, 1–24.

modENCODE Consortium, T., Roy, S., Ernst, J., Kharchenko, P. V, Kheradpour, P., Negre, N., Eaton, M.L., Landolin, J.M., Bristow, C.A., Ma, L., et al. (2011). Identification of Functional Elements and Regulatory Circuits by Drosophila modENCODE. Science (80-.). *330*, 1787–1797.

Nightingale, K.P., O'Neill, L.P., and Turner, B.M. (2006). Histone modifications: Signalling receptors and potential elements of a heritable epigenetic code. Curr. Opin. Genet. Dev. *16*, 125–136.

Niopek, D., Benzinger, D., Roensch, J., Draebing, T., Wehler, P., Eils, R., and Di Ventura, B. (2014). Engineering light-inducible nuclear localization signals for precise spatiotemporal control of protein dynamics in living cells. Nat. Commun. *5*.

Otterstrom, J., Castells-Garcia, A., Vicario, C., Gomez-Garcia, P.A., Cosma, M.P., and Lakadamyali, M. (2019). Super-resolution microscopy reveals how histone tail acetylation affects DNA compaction within nucleosomes in vivo. Nucleic Acids Res. *47*, 8470–8484.

Özel, M.N., Simon, F., Jafari, S., Holguera, I., Chen, Y.C., Benhra, N., El-Danaf, R.N., Kapuralin, K., Malin, J.A., Konstantinides, N., et al. (2021). Neuronal diversity and convergence in a visual system developmental atlas. Nature *589*, 88–95.

Park, P.J. (2009). ChIP-seq: advantages and challenges of a maturing technology. Nat. Rev. Genet. *10*, 669–680.

Pierce, S.E., Granja, J.M., and Greenleaf, W.J. (2021). High-throughput single-cell chromatin accessibility CRISPR screens enable unbiased identification of regulatory networks in cancer. Nat. Commun. *12*, 1–8.

Pijuan-Sala, B., Griffiths, J.A., Guibentif, C., Hiscock, T.W., Jawaid, W., Calero-Nieto, F.J., Mulas, C., Ibarra-Soria, X., Tyser, R.C.V., Ho, D.L.L., et al. (2019). A single-cell molecular map of mouse gastrulation and early organogenesis. Nature *566*, 490–495.

Pijuan-Sala, B., Wilson, N.K., Xia, J., Hou, X., Hannah, R.L., Kinston, S., Calero-Nieto, F.J., Poirion, O., Preissl, S., Liu, F., et al. (2020). Single-cell chromatin accessibility maps reveal regulatory programs driving early mouse organogenesis. Nat. Cell Biol. *22*, 487–497.

Plass, M., Solana, J., Alexander Wolf, F., Ayoub, S., Misios, A., Glažar, P., Obermayer, B., Theis, F.J., Kocks, C., and Rajewsky, N. (2018). Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. Science (80-.). *360*.

Pliner, H.A., Packer, J.S., McFaline-Figueroa, J.L., Cusanovich, D.A., Daza, R.M., Aghamirzaie, D., Srivatsan, S., Qiu, X., Jackson, D., Minkina, A., et al. (2018). Cicero Predicts cis -Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. Mol. Cell 1–14.

Pott, S. (2016). Simultaneous measurement of chromatin accessibility, DNA methylation, and nucleosome phasing in single cells. Elife 061739.

Pott, S., and Lieb, J.D. (2015). Single-cell ATAC-seq: strength in numbers. Genome Biol. *16*, 172.

Preissl, S., Fang, R., Huang, H., Zhao, Y., Raviram, R., Gorkin, D.U., Zhang, Y., Sos, B.C., Afzal, V., Dickel, D.E., et al. (2018). Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. Nat. Neurosci. *21*, 432–439.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Ranzoni, A.M., Tangherloni, A., Berest, I., Riva, S.G., Myers, B., Strzelecka, P.M., Xu, J., Panada, E., Mohorianu, I., Zaugg, J.B., et al. (2021). Integrative Single-Cell RNA-Seq and ATAC-Seq Analysis of Human Developmental Hematopoiesis. Cell Stem Cell *28*, 472-487.

Reddington, J.P., Garfield, D.A., Sigalova, O.M., Karabacak Calviello, A., Marco-Ferreres, R., Girardot, C., Viales, R.R., Degner, J.F., Ohler, U., and Furlong, E.E.M. (2020). Lineage-Resolved Enhancer and Promoter Usage during a Time Course of Embryogenesis. Dev. Cell *55*, 648–664.

Reim, I., and Frasch, M. (2010). Genetic and genomic dissection of cardiogenesis in the drosophila model. Pediatr. Cardiol. *31*, 325–334.

Reiter, F., Wienerroither, S., and Stark, A. (2017). Combinatorial function of transcription factors and cofactors. Curr. Opin. Genet. Dev. *43*, 73–81.

Reznikoff, W.S. (2003). Tn5 as a model for understanding dna transposition. Mol. Microbiol. *47*, 1199–1206.

Reznikoff, W.S. (2008). Transposon Tn5. Annu. Rev. Genet. *42*, 269–286.

Riechmann, V., Irion, U., Wilson, R., Grosskortenhaus, R., and Leptin, M. (1997). Control of cell fates and segmentation in the Drosophila mesoderm. Development *124*, 2915–2922.

Rivera, J., Keränen, S.V.E., Gallo, S.M., and Halfon, M.S. (2019). REDfly: The transcriptional regulatory element database for Drosophila. Nucleic Acids Res. *47*, D828–D834.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–329.

Roman, T.S., Cannon, M.E., Vadlamudi, S., Buchkovich, M.L., Wolford, B.N., Welch, R.P., Morken, M.A., Kwon, G.J., Varshney, A., Kursawe, R., et al. (2017). A Type 2 Diabetes – Associated Functional Regulatory Variant in a Pancreatic Islet Enhancer at the ADCY5 Locus. Diabetes *66*, 2521–2530.

Rossi, M.J., Lai, W.K.M., and Pugh, B.F. (2018). Simplified ChIP-exo assays. Nat. Commun. *9*, 1–13.

Rousseeuw, P.J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. *20*, 53–65.

Rubin, A.J., Parker, K.R., Satpathy, A.T., Qi, Y., Wu, B., Ong, A.J., Mumbach, M.R., Ji, A.L., Kim, D.S., Cho, S.W., et al. (2019). Coupled Single-Cell CRISPR Screening and Epigenomic Profiling Reveals Causal Gene Regulatory Networks. Cell *176*, 361-376.

Rudolf, A., Buttgereit, D., Jacobs, M., Wolfstetter, G., Kesper, D., Pütz, M., Berger, S., Renkawitz-Pohl, R., Holz, A., and Önel, S.F. (2014). Distinct genetic programs guide Drosophila circular and longitudinal visceral myoblast fusion. BMC Cell Biol. *15*.

Ruiz-Gómez, M. (1998). Muscle patterning and specification in Drosophila. Int. J. Dev. Biol. *42*, 283–290.

San Martin, B., and Bate, M. (2001). Hindgut visceral mesoderm requires an ectodermal template for normal development in Drosophila. Development *128*, 233–242.

Sandmann, T., Jensen, L.J., Jakobsen, J.S., Karzynski, M.M., Eichenlaub, M.P., Bork, P., and Furlong, E.E.M. (2006). A Temporal Map of Transcription Factor Activity: Mef2 Directly Regulates Target Genes at All Stages of Muscle Development. Dev. Cell *10*, 797–807.

Sandmann, T., Girardot, C., Brehme, M., Tongprasit, W., Stolc, V., and Furlong, E.E.M. (2007). A core transcriptional network for early mesoderm development in Drosophila melanogaster. Genes Dev. *21*, 436–449.

Satpathy, A.T., Granja, J.M., Yost, K.E., Qi, Y., Meschi, F., McDermott, G.P., Olsen, B.N., Mumbach, M.R., Pierce, S.E., Corces, M.R., et al. (2019). Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. Nat. Biotechnol. *37*, 925–936.

Schaub, C., and Frasch, M. (2013). Org-1 is required for the diversification of circular visceral muscle founder cells and normal midgut morphogenesis. Dev. Biol. *376*, 245–259.

Schep, A.N., Wu, B., Buenrostro, J.D., and Greenleaf, W.J. (2017). ChromVAR: Inferring transcription-factor-associated accessibility from single-cell epigenomic data. Nat. Methods *14*, 975–978.

Schor, I.E., Degner, J.F., Harnett, D., Cannavò, E., Casale, F.P., Shim, H., Garfield, D.A., Birney, E., Stephens, M., Stegle, O., et al. (2017). Promoter shape varies across populations and affects promoter evolution and expression noise. Nat. Genet. *49*, 550–558.

Shashikant, T., Khor, J.M., and Ettensohn, C.A. (2018). Global analysis of primary mesenchyme cell cis-regulatory modules by chromatin accessibility profiling. BMC Genomics *19*, 1–18.

Shlyueva, D., Stampfel, G., and Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. Nat. Rev. Genet. *15*, 272–286.

Sinnamon, J.R., Torkenczy, K.A., Linhoff, M.W., Vitak, S.A., Mulqueen, R.M., Pliner, H.A., Trapnell, C., Steemers, F.J., Mandel, G., and Adey, A.C. (2019). The accessible chromatin landscape of the murine hippocampus at single-cell resolution. Genome Res. *29*, 857–869.

Soshnikova, N., and Duboule, D. (2009). Epigenetic Temporal Control of Mouse Hox Genes in Vivo. Science (80-.). *324*, 1320–1323.

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

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

Sun, S., Zhu, J., Ma, Y., and Zhou, X. (2019). Accuracy, Robustness and Scalability of Dimensionality Reduction Methods for Single Cell RNAseq Analysis. Genome Biol. *20*.

Sun, W., Poschmann, J., Cruz-Herrera del Rosario, R., Parikshak, N.N., Hajan, H.S., Kumar, V., Ramasamy, R., Belgard, T.G., Elanggovan, B., Wong, C.C.Y., et al. (2016). Histone Acetylome-wide Association Study of Autism Spectrum Disorder. Cell *167*, 1385–1397.

Tao, Y., Wang, J., Tokusumi, T., Gajewski, K., and Schulz, R.A. (2007). Requirement of the LIM Homeodomain Transcription Factor Tailup for Normal Heart and Hematopoietic Organ Formation in Drosophila melanogaster. Mol. Cell. Biol. *27*, 3962–3969.

Thomas, S., Li, X.Y., Sabo, P.J., Sandstrom, R., Thurman, R.E., Canfield, T.K., Giste, E., Fisher, W., Hammonds, A., Celniker, S.E., et al. (2011). Dynamic reprogramming of chromatin accessibility during Drosophila embryo development. Genome Biol. *12*.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75–82.

Tintori, S.C., Osborne Nishimura, E., Golden, P., Lieb, J.D., and Goldstein, B. (2016). A Transcriptional Lineage of the Early C. elegans Embryo. Dev. Cell *38*, 430–444.

Tomancak, P., Beaton, A., Weiszmann, R., Kwan, E., Shu, S.Q., Lewis, S.E., Richards, S., Ashburner, M., Hartenstein, V., Celniker, S.E., et al. (2002). Systematic determination of patterns of gene expression during Drosophila embryogenesis. Genome Biol. *3*, 1–14.

Tyser, R.C.V., Ibarra-Soria, X., McDole, K., Jayaram, S.A., Godwin, J., Brand, T.A.H.V. Den, Miranda, A.M.A., Scialdone, A., Keller, P.J., Marioni, J.C., et al. (2021). Characterization of a common progenitor pool of the epicardium and myocardium. Science (80-.). *371*.

Uyehara, C.M., Nystrom, S.L., Niederhuber, M.J., Leatham-Jensen, M., Ma, Y., Buttitta, L.A., and McKay, D.J. (2017). Hormone-dependent control of developmental timing through regulation of chromatin accessibility. Genes Dev. *31*, 862–875.

Wang, Q., Xiong, H., Ai, S., Yu, X., Liu, Y., Zhang, J., and He, A. (2019). CoBATCH for High-Throughput Single-Cell Epigenomic Profiling. Mol. Cell *76*, 206–216.

Weintraub, H., Tapscott, S.J., Davis, R.L., Thayer, M.J., Adam, M. a, Lassar, A.B., and Miller, a D. (1989). Activation of muscle-specific genes in pigment, nerve, fat, liver, and fibroblast cell lines by forced expression of MyoD. Dev. Biol. *86*, 5434–5438.

Wilczyski, B., and Furlong, E.E.M. (2010). Dynamic CRM occupancy reflects a temporal map of developmental progression. Mol. Syst. Biol. *6*.

Wilmut, I., Schnieke, A.E., McWhir, J., Kind, A.J., and Campbell, K.H.S. (1997). Viable offspring derived from fetal and adult mammalian cells. Nature *385*, 810–813.

Wittkopp, P.J., and Kalay, G. (2012). Cis-regulatory elements: Molecular mechanisms and evolutionary processes underlying divergence. Nat. Rev. Genet. *13*, 59–69.

Yan, W., Chen, D., Schumacher, J., Durantini, D., Engelhorn, J., Chen, M., Carles, C.C., and Kaufmann, K. (2019). Dynamic control of enhancer activity drives stage-specific gene expression during flower morphogenesis. Nat. Commun. *10*, 1–16.

Yin, Z., and Frasch, M. (1998). Regulation and function of tinman during dorsal mesoderm induction and heart specification in Drosophila. Dev. Genet. *22*, 187–200.

Yin, Z., Xu, X.L., and Frasch, M. (1997). Regulation of the twist target gene tinman by modular cis-regulatory elements during early mesoderm development. Development *124*, 4971–4982.

Zaffran, S., Küchler, A., Lee, H.H., and Frasch, M. (2001). biniou (FoxF), a central component in a regulatory network controlling visceral mesoderm development and midgut morphogenesis in Drosophila. Genes Dev. *15*, 2900–2915.

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

Zinzen, R.P., Girardot, C., Gagneur, J., Braun, M., and Furlong, E.E.M. (2009). Combinatorial binding predicts spatio-temporal cis-regulatory activity. Nature *462*, 65–70.

Zmojdzian, M., and Jagla, K. (2013). Tailup plays multiple roles during cardiac outflow assembly in Drosophila. Cell Tissue Res. *354*, 639–645.

I hereby declare that I, Stefano Secchia, prepared this Ph.D. thesis:

'Single-cell dissection of regulatory landscapes during embryogenesis'

on my own and with no other sources and aids than quoted.

8 Data Tables

Data Table 1. Top 100 DA genes by myogenic population.

Flybase ID	Gene name	Log2 fold-change	Annotation
FBgn0002561	l(1)sc	1.64814028759505	Mesoderm
FBgn0040487	BobA	1.5342779157903	Mesoderm
FBgn0262398	mir-124	1.44283706121715	Mesoderm
FBgn0262397	mir-2c	1.38788469682677	Mesoderm
FBgn0034224	insb	1.36918715006522	Mesoderm
FBgn0053200	VepD	1.32792375069709	Mesoderm
FBgn0262411	mir-92a	1.27996717029096	Mesoderm
FBgn0000216	Brd	1.27975709400999	Mesoderm
FBgn0262462	mir-iab-4	1.27673361281348	Mesoderm
FBgn0040809	CG13465	1.27621230029935	Mesoderm
FBgn0013725	phyl	1.20269269704021	Mesoderm
FBgn0262429	mir-13a	1.18681786853675	Mesoderm
FBgn0002631	E(spl)m5-HLH	1.17433956718042	Mesoderm
FBgn0035643	CG13287	1.11107729568015	Mesoderm
FBgn0003720	tll	1.07727523974007	Mesoderm
FBgn0262367	mir-13b-1	1.04471291550831	Mesoderm
FBgn0001174	halo	1.0429291737583	Mesoderm
FBgn0002629	E(spl)m4-BFM	1.04292426366589	Mesoderm
FBgn0000180	bib	1.03723484104608	Mesoderm
FBgn0069056	CG33226	1.02471610089666	Mesoderm
FBgn0010382	CycE	1.02135397453868	Mesoderm
FBgn0027364	Six4	0.999383604901696	Mesoderm
FBgn0030477	dmrt11E	0.993144679999179	Mesoderm
FBgn0086677	jeb	0.97574535375905	Mesoderm
FBgn0032493	Mabi	0.9738284079205	Mesoderm
FBgn0040743	CG15919	0.966920332630115	Mesoderm
FBgn0040385	CG12496	0.958271783160701	Mesoderm
FBgn0000591	E(spl)m8-HLH	0.954278441150445	Mesoderm
FBgn0262461	mir-315	0.95038729151486	Mesoderm
FBgn0026320	Tom	0.949976167397672	Mesoderm
FBgn0001077	ftz	0.949827230392773	Mesoderm
FBgn0031757	Ucp4C	0.943863467754309	Mesoderm
FBgn0039286	dan	0.933330784681837	Mesoderm
FBgn0083946	lobo	0.930019478631484	Mesoderm
FBgn0053182	Kdm4B	0.924898634224354	Mesoderm
FBgn0040296	Ocho	0.902829807096362	Mesoderm
FBgn0261434	hkb	0.89711570129262	Mesoderm

FBgn0028991	seq	0.896876609286206	Mesoderm
FBgn0262370	mir-7	0.895614059508022	Mesoderm
FBgn0033802	CG17724	0.891831641246368	Mesoderm
FBgn0002632	E(spl)m6-BFM	0.88845731622931	Mesoderm
FBgn0034158	CG5522	0.865477763781129	Mesoderm
FBgn0000463	Dl	0.859279867789802	Mesoderm
FBgn0263042	CG43337	0.850444346287723	Mesoderm
FBgn0032491	CG16815	0.849655242764447	Mesoderm
FBgn0038134	wntD	0.849169552201324	Mesoderm
FBgn0262839	CG43203	0.844240746654488	Mesoderm
FBgn0050287	CG30287	0.84270715194591	Mesoderm
FBgn0024249	cato	0.842561244022952	Mesoderm
FBgn0000137	ase	0.839316828165909	Mesoderm
FBgn0025776	ind	0.832931452234417	Mesoderm
FBgn0004896	fd59A	0.831943163624262	Mesoderm
FBgn0003731	Egfr	0.826863223416677	Mesoderm
FBgn0038833	CG15696	0.826221803459343	Mesoderm
FBgn0261723	Dbx	0.819925001135757	Mesoderm
FBgn0029740	CG12680	0.819426752864759	Mesoderm
FBgn0265632	CG44439	0.818994869822472	Mesoderm
FBgn0085305	CG34276	0.814698250643091	Mesoderm
FBgn0267791	HnRNP-K	0.81010782078488	Mesoderm
FBgn0040992	CG10570	0.80045993518551	Mesoderm
FBgn0015371	chn	0.80032685059186	Mesoderm
FBgn0003607	Su(var)205	0.799404924458442	Mesoderm
FBgn0033458	Lime	0.796556926408278	Mesoderm
FBgn0262876	CG43231	0.794904274605054	Mesoderm
FBgn0004795	retn	0.782413299578101	Mesoderm
FBgn0021776	mira	0.777610478735535	Mesoderm
FBgn0031999	CG8419	0.773810629346246	Mesoderm
FBgn0004173	Mst84Db	0.772983631183112	Mesoderm
FBgn0262365	CG43063	0.76951819560449	Mesoderm
FBgn0259789	zld	0.747817723591224	Mesoderm
FBgn0053557	CG33557	0.746780414402127	Mesoderm
FBgn0259745	wech	0.734225205664695	Mesoderm
FBgn0086675	fne	0.725432369268138	Mesoderm
FBgn0036503	CG13454	0.724202881636184	Mesoderm
FBgn0004172	Mst84Da	0.720479834830503	Mesoderm
FBgn0037562	Nazo	0.706688282876428	Mesoderm
FBgn0008636	hbn	0.701055581182874	Mesoderm
FBgn0035558	CG11357	0.698239006411993	Mesoderm
FBgn0000320	eya	0.697977925920366	Mesoderm

FBgn0028513	CG9254	0.694937611033274	Mesoderm
FBgn0260477	CG30283	0.69385620665698	Mesoderm
FBgn0262449	mir-2b-2	0.687233742009687	Mesoderm
FBgn0003430	slp1	0.684835133936007	Mesoderm
FBgn0036494	Toll-6	0.684509600646735	Mesoderm
FBgn0030510	CG12177	0.683802278172372	Mesoderm
FBgn0036345	CG17300	0.683531535936297	Mesoderm
FBgn0000071	Ama	0.68103590601958	Mesoderm
FBgn0038385	Fbxl7	0.680374161983235	Mesoderm
FBgn0002973	numb	0.679357668474015	Mesoderm
FBgn0052282	Drsl4	0.678581711631486	Mesoderm
FBgn0050398	CG30398	0.670118703754767	Mesoderm
FBgn0015609	CadN	0.668282960316334	Mesoderm
FBgn0039523	CG12885	0.667698580154408	Mesoderm
FBgn0039288	CG13653	0.666224056160738	Mesoderm
FBgn0262290	mir-iab-8	0.656280974551549	Mesoderm
FBgn0010109	dpn	0.652462615122235	Mesoderm
FBgn0000996	dup	0.651143878675945	Mesoderm
FBgn0262377	mir-2a-1	0.647799656398593	Mesoderm
FBgn0004110	tin	0.647782909725355	Mesoderm
FBgn0004647	N	0.644815636453689	Mesoderm
FBgn0262459	mir-34	1.4642442160037	Somatic
FBgn0050281	CG30281	1.19540348758045	Somatic
FBgn0264442	ab	1.14325053884973	Somatic
FBgn0050280	CG30280	1.13870942779105	Somatic
FBgn0037447	Neurochondrin	1.1293114119408	Somatic
FBgn0263118	tx	1.11763936781541	Somatic
FBgn0243516	Vrp1	1.10522117631259	Somatic
FBgn0033782	sug	1.09678634453405	Somatic
FBgn0010423	TpnC47D	1.09305223580861	Somatic
FBgn0002772	Mlc1	1.08684114367607	Somatic
FBgn0033781	CG13319	1.07292392768228	Somatic
FBgn0030073	CG10962	1.03548373103088	Somatic
FBgn0039007	CCAP	1.01356989242095	Somatic
FBgn0031313	CG5080	1.00075967548389	Somatic
FBgn0050350	CG30350	0.988336460714625	Somatic
FBgn0029648	CG3603	0.988143954228553	Somatic
FBgn0036663	CG9674	0.982553534461279	Somatic
FBgn0016694	Pdp1	0.968686133801794	Somatic
FBgn0000116	Argk	0.966112806077584	Somatic
FBgn0265180	CG44245	0.960336977287001	Somatic
FBgn0040600	CG13631	0.957686528642686	Somatic

FBgn0031908	CG5177	0.956322260326828	Somatic
FBgn0039756	CG9743	0.953636318445761	Somatic
FBgn0039755	CG15531	0.95295462097472	Somatic
FBgn0262368	mir-282	0.951615185353277	Somatic
FBgn0265191	Gyg	0.945583800273223	Somatic
FBgn0028841	jhamt	0.923200482623471	Somatic
FBgn0262594	CheA46a	0.916194667317664	Somatic
FBgn0042711	Hex-t1	0.907927172704414	Somatic
FBgn0053548	SmydA-8	0.895988516727152	Somatic
FBgn0036030	Prps	0.892688052785885	Somatic
FBgn0262447	mir-14	0.891865124274097	Somatic
FBgn0001258	Ldh	0.887953185004414	Somatic
FBgn0259973	Sfp79B	0.879854951920239	Somatic
FBgn0010424	TpnC73F	0.879184041050892	Somatic
FBgn0027073	Ugt49B1	0.872671615457961	Somatic
FBgn0259209	Mlp60A	0.866388737446798	Somatic
FBgn0267486	Ptp36E	0.866219358170787	Somatic
FBgn0020445	E23	0.851641591084222	Somatic
FBgn0036152	CG6175	0.850763507563695	Somatic
FBgn0024321	NK7.1	0.848056754817383	Somatic
FBgn0005638	slbo	0.847158128361925	Somatic
FBgn0004169	up	0.839830162842417	Somatic
FBgn0284237	CG46321	0.837972228172331	Somatic
FBgn0035924	CG6576	0.837418270865528	Somatic
FBgn0259701	CG42355	0.828096253867383	Somatic
FBgn0039008	CG6972	0.824034067694867	Somatic
FBgn0052813	CG32813	0.821022008498574	Somatic
FBgn0263546	mir-4985	0.820719157629318	Somatic
FBgn0263117	CG34377	0.818917399245946	Somatic
FBgn0003888	betaTub60D	0.814996028884772	Somatic
FBgn0030887	CG6867	0.813512194986806	Somatic
FBgn0265487	mbl	0.803794196566757	Somatic
FBgn0001078	ftz-f1	0.797071511087713	Somatic
FBgn0000064	Ald1	0.795675383056507	Somatic
FBgn0002773	Mlc2	0.79302267872658	Somatic
FBgn0040260	Ugt37D1	0.792076257956238	Somatic
FBgn0035313	CG13810	0.790676912291763	Somatic
FBgn0034694	Plekhm1	0.785511658745901	Somatic
FBgn0037956	CG6959	0.781286848968723	Somatic
FBgn0035697	CG10163	0.768604876369528	Somatic
FBgn0265512	mlt	0.765205937072012	Somatic
FBgn0002789	Mp20	0.763897975798888	Somatic

FBgn0033497	CG12912	0.761459965075712	Somatic
FBgn0085235	CG34206	0.757421044455956	Somatic
FBgn0283471	wupA	0.753981988128249	Somatic
FBgn0015321	Ubc4	0.748794313415922	Somatic
FBgn0042185	MCU	0.748745050038663	Somatic
FBgn0086355	Tpi	0.74684735638844	Somatic
FBgn0037144	CG7458	0.746603296392902	Somatic
FBgn0263973	jv	0.745707004973689	Somatic
FBgn0004133	blow	0.744454216202896	Somatic
FBgn0004117	Tm2	0.741968519134169	Somatic
FBgn0035143	Ppm1	0.74176805006609	Somatic
FBgn0033141	CG12831	0.739562069635866	Somatic
FBgn0001225	Hsp26	0.729219751963215	Somatic
FBgn0037835	CG14687	0.726452886715346	Somatic
FBgn0024189	sns	0.722471476107517	Somatic
FBgn0001142	Gs1	0.720754444707982	Somatic
FBgn0000448	Hr3	0.718933380429846	Somatic
FBgn0262855	CG43219	0.718086406757412	Somatic
FBgn0051140	CG31140	0.715812873129707	Somatic
FBgn0026263	bip1	0.713575958182013	Somatic
FBgn0265356	tn	0.711114765808807	Somatic
FBgn0035014	CG13581	0.707019947944414	Somatic
FBgn0039419	CG12290	0.706169856904387	Somatic
FBgn0032400	CG6770	0.705355165925004	Somatic
FBgn0000567	Eip74EF	0.702313386974563	Somatic
FBgn0261549	rdgA	0.700466792048632	Somatic
FBgn0033153	Gadd45	0.695018301426055	Somatic
FBgn0050359	Mal-A5	0.693612053685483	Somatic
FBgn0035798	frac	0.693179243303302	Somatic
FBgn0042710	Hex-t2	0.6889794142826	Somatic
FBgn0039719	CG15515	0.682240302443935	Somatic
FBgn0036377	CG10710	0.681064128359075	Somatic
FBgn0032129	jp	0.679392617535073	Somatic
FBgn0004865	Eip78C	0.678750604008161	Somatic
FBgn0029518	CG13376	0.678660133876374	Somatic
FBgn0261836	Msp300	0.673065875429426	Somatic
FBgn0052369	CG32369	0.666485639532858	Somatic
FBgn0003117	pnr	1.35707723990039	Cardiac
FBgn0260761	CG42559	1.09270782963942	Cardiac
FBgn0262323	mir-252	0.971043904405741	Cardiac
FBgn0262020	CG42831	0.931749545739967	Cardiac
FBgn0263321	CG43402	0.859149500006809	Cardiac

FBgn0036262	CG6910	0.821909219908768	Cardiac
FBgn0035954	Doc3	0.819532630190535	Cardiac
FBgn0050272	MFS1	0.809975586958541	Cardiac
FBgn0051821	CG31821	0.79016038689697	Cardiac
FBgn0031959	spz3	0.737389958338737	Cardiac
FBgn0033866	CG6280	0.733309260165377	Cardiac
FBgn0003896	tup	0.717574287660192	Cardiac
FBgn0015903	apt	0.712018513690028	Cardiac
FBgn0260003	Dys	0.677373879973993	Cardiac
FBgn0025641	DAAM	0.653760490761615	Cardiac
FBgn0266801	CG45263	0.637554524050994	Cardiac
FBgn0039736	CG7912	0.618379404485206	Cardiac
FBgn0011591	fng	0.600018901151866	Cardiac
FBgn0262475	bru2	0.586630260614513	Cardiac
FBgn0038498	beat-IIa	0.578323247702471	Cardiac
FBgn0260440	spdo	0.574261909614073	Cardiac
FBgn0002632	E(spl)m6-BFM	0.534652600774903	Cardiac
FBgn0002734	HLH	0.532133597952217	Cardiac
FBgn0086613	Ino80	0.529358797752515	Cardiac
FBgn0038391	GATAe	0.519790382426994	Cardiac
FBgn0263392	Tet	0.500051722354527	Cardiac
FBgn0037114	Cpr78E	1.92196153879039	Visceral
FBgn0262417	mir-316	1.25971169725524	Visceral
FBgn0262415	mir-289	1.25312278583402	Visceral
FBgn0000636	Fas3	1.23338246953143	Visceral
FBgn0085200	CG34171	1.18566507042087	Visceral
FBgn0283551	mir-9378	1.09547316284342	Visceral
FBgn0001170	H2.0	1.05992360085057	Visceral
FBgn0000244	by	1.01430165943397	Visceral
FBgn0039075	CG4393	0.946222631593539	Visceral
FBgn0264001	bru3	0.936448388860649	Visceral
FBgn0011559	Acp36DE	0.935644159281751	Visceral
FBgn0263930	dally	0.933233412097112	Visceral
FBgn0037115	CG11249	0.921023105767852	Visceral
FBgn0005564	Shal	0.904275834543947	Visceral
FBgn0010015	CanA1	0.891638009113778	Visceral
FBgn0030796	CG4829	0.860619126346195	Visceral
FBgn0038912	CG6656	0.815378075698644	Visceral
FBgn0032297	CG17124	0.803642342356888	Visceral
FBgn0262788	CG43169	0.803004861252479	Visceral
FBgn0045759	bin	0.791447283197346	Visceral
FBgn0015773	NetA	0.788737991438492	Visceral

FBgn0004862	bap	0.75763559093031	Visceral
FBgn0000473	Cyp6a2	0.74676752741692	Visceral
FBgn0035144	Kah	0.746611614680719	Visceral
FBgn0033149	CG11060	0.742752639694124	Visceral
FBgn0040699	CG15024	0.742734634910444	Visceral
FBgn0032209	Hand	0.740233563158251	Visceral
FBgn0262327	mir-2281	0.737027328637424	Visceral
FBgn0037835	CG14687	0.727438581773945	Visceral
FBgn0035711	CG8519	0.724488014026316	Visceral
FBgn0046332	gskt	0.722172084503719	Visceral
FBgn0025631	moody	0.722157215843978	Visceral
FBgn0262842	CG43206	0.717876385209913	Visceral
FBgn0052026	CG32026	0.714179514720203	Visceral
FBgn0085376	CG34347	0.712813034164892	Visceral
FBgn0261648	salm	0.710014315989274	Visceral
FBgn0036121	CG6310	0.7068437838427	Visceral
FBgn0262534	CG43088	0.70598772995349	Visceral
FBgn0050385	CG30385	0.699339199414236	Visceral
FBgn0010423	TpnC47D	0.69495484129827	Visceral
FBgn0035085	CG3770	0.694690757730581	Visceral
FBgn0050384	CG30384	0.693111710147623	Visceral
FBgn0085409	smal	0.684912621113465	Visceral
FBgn0031972	Wwox	0.684909251334474	Visceral
FBgn0264711	CG43980	0.683767551619037	Visceral
FBgn0052713	CG32713	0.679543117200088	Visceral
FBgn0034638	CG10433	0.674754337393292	Visceral
FBgn0010909	msn	0.673653126545629	Visceral
FBgn0262594	CheA46a	0.671239249494161	Visceral
FBgn0016930	Dyrk2	0.665686650058557	Visceral
FBgn0250849	CG32388	0.660036754906033	Visceral
FBgn0053140	CG33140	0.647850904857036	Visceral
FBgn0044011	Spn43Ad	0.641753709673575	Visceral
FBgn0024846	p38b	0.641190426373274	Visceral
FBgn0030797	CG13004	0.637936286122249	Visceral
FBgn0003090	pk	0.637115883105075	Visceral
FBgn0032587	CG5953	0.63376281633157	Visceral
FBgn0038156	side-IV	0.632329335095486	Visceral
FBgn0050156	CG30156	0.632184389178929	Visceral
FBgn0085421	Epac	0.629016290599202	Visceral
FBgn0002930	nec	0.628217540754661	Visceral
FBgn0025632	CG4313	0.62498589229514	Visceral
FBgn0035084	CG15861	0.62173895036904	Visceral

FBgn0034477	CG13872	0.613525110141341	Visceral
FBgn0261560	Thor	0.612183922751963	Visceral
FBgn0036259	CG9760	0.610097350248777	Visceral
FBgn0035542	DOR	0.609495804964936	Visceral
FBgn0024293	Spn43Ab	0.604281841876772	Visceral
FBgn0040765	luna	0.604031809298155	Visceral
FBgn0037007	BNIP3	0.603494252572838	Visceral
FBgn0020280	laf	0.597083179122445	Visceral
FBgn0030992	CG33253	0.590525317355336	Visceral
FBgn0050421	Usp15-31	0.587820959808631	Visceral
FBgn0039500	CG5984	0.587007094167366	Visceral
FBgn0037448	CG15186	0.579812386712392	Visceral
FBgn0031747	CG9021	0.578643266062929	Visceral
FBgn0035143	Ppm1	0.571984067034331	Visceral
FBgn0036474	Or71a	0.571249279887949	Visceral
FBgn0031792	CG13983	0.569834731782572	Visceral
FBgn0024294	Spn43Aa	0.568082310176352	Visceral
FBgn0036242	CG6793	0.565630028605625	Visceral
FBgn0026376	Rgl	0.56522808025499	Visceral
FBgn0263555	mir-4943	0.557952233844876	Visceral
FBgn0037116	Als2	0.556959488971591	Visceral
FBgn0005775	Con	0.555274973719536	Visceral
FBgn0033677	CG8321	0.550714508999279	Visceral
FBgn0041092	tai	0.549547006102324	Visceral
FBgn0264725	CG43993	0.546855479608547	Visceral
FBgn0022960	vimar	0.545717583073847	Visceral
FBgn0033524	Cyp49a1	0.5456597605398	Visceral
FBgn0037623	CG9801	0.542226072471667	Visceral
FBgn0263118	tx	0.5410469278445	Visceral
FBgn0040505	Alk	0.540395441823949	Visceral
FBgn0033145	CG12828	0.536828771451337	Visceral
FBgn0036196	CG11658	0.535347408236176	Visceral
FBgn0039178	CG6356	0.534182032643054	Visceral
FBgn0036141	wls	0.533348800049293	Visceral
FBgn0265140	Meltrin	0.529577050242896	Visceral
FBgn0051816	CG31816	0.527654234220256	Visceral
FBgn0021767	org-1	0.527372741131558	Visceral

Data Table 2. Differentially accessible regions identified in the *Mef2* mutant.

Region coordinates	Group	Muscle enhancer	Closest gene
chr3L:22787983-22788318	Mef2-gain	NA	CG14448
chr2R:9543777-9544417	Mef2-gain	NA	CG1888
chr2R:6524445-6524941	Mef2-gain	NA	jing
chr3R:26796545-26796972	Mef2-gain	NA	Tl
chr3L:16641761-16642182	Mef2-loss	NA	Abl
chr2R:20946563-20946863	Mef2-loss	NA	Act57B
chr2R:12595492-12595792	Mef2-loss	FALSE	Amph
chr3R:17346246-17346600	Mef2-loss	NA	beat-IIa
chr3R:17386619-17386919	Mef2-loss	NA	Brf
chr2L:16231212-16231722	Mef2-loss	NA	Ca-Ma2d
chr3R:20990364-20990714	Mef2-loss	NA	Calx
chr2R:12273104-12273552	Mef2-loss	NA	Cam
chr2L:11272404-11272704	Mef2-loss	NA	cana
chr3R:22701517-22702159	Mef2-loss	NA	CCAP
chr3L:12836325-12836815	Mef2-loss	NA	CG10960
chr3L:11703590-11704020	Mef2-loss	NA	CG11658
chr3R:4780178-4780485	Mef2-loss	NA	CG14655
chr3R:6481462-6482091	Mef2-loss	NA	CG15186
chr3R:21566368-21566958	Mef2-loss	NA	CG15498
chr3L:2155660-2155966	Mef2-loss	NA	CG15822
chr3R:30175167-30175467	Mef2-loss	TRUE	CG1983
chr2R:17077833-17078494	Mef2-loss	NA	CG30460
chr3R:13543039-13543604	Mef2-loss	NA	CG31337
chr2L:2731772-2732270	Mef2-loss	NA	CG31689
chr3L:3788506-3788834	Mef2-loss	NA	CG32264
chr2L:16874923-16875379	Mef2-loss	NA	CG32832
chr3L:10783510-10783818	Mef2-loss	NA	CG43245
chr2L:3765801-3766201	Mef2-loss	NA	CG43707
chr3R:8684803-8685362	Mef2-loss	NA	CG44227
chr3R:10777230-10777616	Mef2-loss	NA	CG45076
chr2L:1162156-1162632	Mef2-loss	TRUE	CG5080
chr2L:7405035-7405360	Mef2-loss	NA	CG5177
chr2L:7407403-7408132	Mef2-loss	NA	CG5177
chr3R:27247514-27247814	Mef2-loss	NA	CG5984
chr3R:19107852-19108558	Mef2-loss	NA	CG6040
chr3L:15178431-15179021	Mef2-loss	NA	CG7011
chr3R:22784545-22785230	Mef2-loss	NA	CG7029
chr3L:15675233-15675600	Mef2-loss	NA	CG7304
chr3L:11424291-11424739	Mef2-loss	NA	CG7560
chr2R:9749781-9750495	Mef2-loss	NA	CheA46a

chr3L:8849533-8850412	Mef2-loss	NA	dally
chr3R:6393911-6394764	Mef2-loss	NA	Dmtn
chr3R:28910691-28911593	Mef2-loss	NA	Doa
chr3R:19486035-19486551	Mef2-loss	NA	Dys
chr3R:29496251-29496633	Mef2-loss	NA	FipoQ
chr3L:13413511-13413819	Mef2-loss	NA	flr
chr3L:7562458-7562767	Mef2-loss	NA	frac
chr3L:1235647-1236230	Mef2-loss	NA	galene
chr3R:6409770-6410070	Mef2-loss	NA	gpp
chr3R:19002514-19002954	Mef2-loss	NA	gukh
chr3L:8676999-8677299	Mef2-loss	TRUE	h
chr2R:17757079-17757444	Mef2-loss	TRUE	HLH54F
chr3R:22047306-22047710	Mef2-loss	NA	how
chr3R:10595474-10595774	Mef2-loss	TRUE	hth
chr3R:10600721-10601423	Mef2-loss	TRUE	hth
chr2R:19411237-19411695	Mef2-loss	NA	hts
chr3R:21616691-21616991	Mef2-loss	NA	InR
chr2L:9557542-9558255	Mef2-loss	NA	jp
chr2L:9551096-9551528	Mef2-loss	NA	jp
chr3L:535272-535887	Mef2-loss	NA	klar
chr3L:16708803-16709240	Mef2-loss	NA	Lasp
chr2R:16162662-16163090	Mef2-loss	NA	lbk
chr2R:16163118-16163418	Mef2-loss	NA	lbk
chr3L:16910980-16911454	Mef2-loss	NA	Lmpt
chr3R:21863267-21863567	Mef2-loss	NA	lsn
chr2L:6771929-6772290	Mef2-loss	NA	LUBEL
chr2R:11104216-11104516	Mef2-loss	NA	luna
chr2R:8462658-8462958	Mef2-loss	NA	Mal-A5
chr3L:591236-591687	Mef2-loss	NA	MED14
chr3L:592329-592785	Mef2-loss	NA	MED14
chr2L:16766377-16766942	Mef2-loss	TRUE	Mhc
chr2L:16768257-16768557	Mef2-loss	TRUE	Mhc
chr2L:16767566-16767866	Mef2-loss	TRUE	Mhc
chr3R:16074122-16074423	Mef2-loss	NA	Mhcl
chr2L:19037131-19037436	Mef2-loss	NA	mib2
chr2L:20484460-20484760	Mef2-loss	NA	mir-1
chr3L:10352274-10352610	Mef2-loss	FALSE	mir-276a
chr3L:634991-635446	Mef2-loss	NA	mir-ban
chr3R:27659115-27659415	Mef2-loss	TRUE	Mlc1
chr3R:30174393-30175135	Mef2-loss	TRUE	Mlc2
chr3R:24608830-24609971	Mef2-loss	NA	mld
chr2R:24074094-24074848	Mef2-loss	NA	Mlp60A

chr2R:13245219-13245583	Mef2-loss	NA	Mp20
chr2R:16211702-16212347	Mef2-loss	NA	mrj
chr3L:2744981-2745387	Mef2-loss	TRUE	Mrtf
chr3L:2748639-2749144	Mef2-loss	NA	Mrtf
chr2L:5101158-5101466	Mef2-loss	NA	Msp300
chr2R:14713424-14713858	Mef2-loss	NA	mspo
chr3R:14332474-14332887	Mef2-loss	NA	NK7.1
chr3R:7658059-7658461	Mef2-loss	NA	Nlg1
chr3R:7661588-7662021	Mef2-loss	NA	Nlg1
chr2L:7711262-7711592	Mef2-loss	NA	Ntl
chr3R:10284170-10284622	Mef2-loss	NA	Nuak1
chr2L:9525544-9525983	Mef2-loss	NA	Oatp30B
chr2L:14685078-14685537	Mef2-loss	NA	osp
chr2R:9861139-9861469	Mef2-loss	NA	PCB
chr3L:7837706-7838083	Mef2-loss	FALSE	Pdp1
chr2L:9706945-9707251	Mef2-loss	NA	pelo
chr3R:23872981-23873422	Mef2-loss	NA	Pli
chr3R:23870863-23871163	Mef2-loss	NA	Pli
chr2L:16761344-16761720	Mef2-loss	NA	ppk17
chr2R:11443863-11444427	Mef2-loss	NA	qvr
chr3L:9841910-9842455	Mef2-loss	NA	RasGAP1
chr2L:5874319-5874701	Mef2-loss	NA	rau
chr2L:2958842-2959269	Mef2-loss	NA	Rbp9
chr2R:22926336-22926670	Mef2-loss	NA	ReepA
chr2R:19321076-19321537	Mef2-loss	NA	Rgk1
chr3L:8553171-8553539	Mef2-loss	NA	rhea
chr3L:294805-295527	Mef2-loss	NA	RhoGEF3
chr2R:19270617-19271432	Mef2-loss	TRUE	rib
chr2L:1723173-1723519	Mef2-loss	NA	Rim2
chr2R:11662684-11662984	Mef2-loss	NA	Roc2
chr2L:20352280-20352701	Mef2-loss	NA	RtGEF
chr2R:15246768-15247344	Mef2-loss	NA	scb
chr2L:19849457-19850119	Mef2-loss	TRUE	sick
chr2L:19864129-19864429	Mef2-loss	NA	sick
chr2L:19847088-19847542	Mef2-loss	NA	sick
chr3R:27402122-27402434	Mef2-loss	NA	side
chr3R:27415863-27416216	Mef2-loss	NA	side
chr3R:13447939-13448521	Mef2-loss	NA	side-IV
chr3R:30110919-30111307	Mef2-loss	NA	sima
chr2R:18442564-18442937	Mef2-loss	NA	slim
chr3L:2087914-2088363	Mef2-loss	NA	sls
chr3L:2089740-2090260	Mef2-loss	TRUE	sls

chr3R:4671307-4671713	Mef2-loss	NA	smash
chr3L:12401702-12402123	Mef2-loss	NA	Smyd4-2
chr2L:15065655-15065986	Mef2-loss	NA	solo
chr3R:14384346-14384646	Mef2-loss	NA	soti
chr2R:9748029-9748523	Mef2-loss	NA	sqa
chr2R:6007196-6007629	Mef2-loss	NA	Src42A
chr2L:18899551-18899957	Mef2-loss	FALSE	ssp3
chr2R:15989297-15989654	Mef2-loss	NA	Strn-Mlck
chr3L:3412016-3412341	Mef2-loss	NA	sty
chr3R:16329765-16330130	Mef2-loss	NA	Sulf1
chr3L:22404069-22404561	Mef2-loss	TRUE	Ten-m
chr2L:6423065-6423605	Mef2-loss	NA	Tig
chr3R:15282315-15282615	Mef2-loss	NA	Tm1
chr2R:18989902-18990258	Mef2-loss	NA	tn
chr2R:18990319-18990706	Mef2-loss	NA	tn
chr2L:3100987-3101318	Mef2-loss	NA	toc
chr2R:11275405-11275763	Mef2-loss	NA	TpnC47D
chr3L:17047698-17047998	Mef2-loss	NA	TpnC73F
chr3L:5754125-5754425	Mef2-loss	NA	Usp47
chr2L:3786951-3787325	nonMef2-gain	NA	bark
chr2L:3789560-3790073	nonMef2-gain	NA	bark
chr3R:4196973-4197438	nonMef2-gain	NA	beta-Man
chr2L:3776772-3777118	nonMef2-gain	NA	bowl
chr2R:24369582-24370061	nonMef2-gain	NA	bs
chr2L:12507628-12508065	nonMef2-gain	FALSE	bun
chr2L:16325767-16326098	nonMef2-gain	NA	cact
chr3R:29841255-29841823	nonMef2-gain	NA	Cad99C
chr2L:13802011-13802552	nonMef2-gain	NA	CAH1
chr2L:18732749-18733323	nonMef2-gain	NA	CG10348
chr3L:12490939-12491239	nonMef2-gain	NA	CG10660
chr3L:12490039-12490887	nonMef2-gain	NA	CG10660
chr2R:8059060-8059439	nonMef2-gain	NA	CG12769
chr2R:15485514-15485814	nonMef2-gain	NA	CG12964
chr3L:4840016-4840418	nonMef2-gain	NA	CG13707
chr3L:1746328-1747099	nonMef2-gain	NA	CG13921
chr2L:15107052-15107795	nonMef2-gain	FALSE	CG15269
chr2L:14456385-14457168	nonMef2-gain	NA	CG15283
chr2R:20000619-20000939	nonMef2-gain	NA	CG16898
chr3L:18996473-18996934	nonMef2-gain	NA	CG18135
chr2L:21669555-21669866	nonMef2-gain	NA	CG2225
chr2L:21672762-21673121	nonMef2-gain	NA	CG2225
chr2L:21954373-21955205	nonMef2-gain	NA	CG2528

chr3R:21721764-21722320	nonMef2-gain	NA	CG31176
chr2L:20845337-20845976	nonMef2-gain	NA	CG31676
chr2L:16859097-16859552	nonMef2-gain	NA	CG31809
chr2R:18664064-18664751	nonMef2-gain	NA	CG33136
chr2L:21029938-21030443	nonMef2-gain	NA	CG42238
chr2L:21009553-21010230	nonMef2-gain	NA	CG42238
chr2L:21026744-21027196	nonMef2-gain	NA	CG42238
chr2L:20960284-20960874	nonMef2-gain	NA	CG42238
chr2L:2242142-2242514	nonMef2-gain	NA	CG4267
chr2R:24808024-24808615	nonMef2-gain	NA	CG42851
chr3R:5109529-5110149	nonMef2-gain	NA	CG43131
chr3L:15532410-15532796	nonMef2-gain	NA	CG43248
chr3L:614828-615571	nonMef2-gain	NA	CG43337
chr3L:7335074-7335426	nonMef2-gain	NA	CG43780
chr3R:19299098-19299399	nonMef2-gain	NA	CG44174
chr3R:19299515-19299857	nonMef2-gain	NA	CG44174
chr3L:16149473-16149811	nonMef2-gain	NA	CG5151
chr2L:9947274-9947798	nonMef2-gain	NA	CG5853
chr3L:11369997-11370446	nonMef2-gain	NA	CG6163
chr3L:18204296-18204840	nonMef2-gain	NA	CG7320
chr3L:18224475-18225034	nonMef2-gain	NA	CG7320
chr2R:25279808-25280159	nonMef2-gain	NA	CG9380
chr2L:13068815-13069115	nonMef2-gain	NA	CG9932
chr2L:2357025-2357808	nonMef2-gain	NA	CG9967
chr3L:18235981-18236342	nonMef2-gain	NA	CheA75a
chr3R:20257192-20257557	nonMef2-gain	NA	cic
chr3R:30817196-30817815	nonMef2-gain	FALSE	cindr
chr2R:19229669-19230145	nonMef2-gain	NA	cora
chr2R:15372937-15373301	nonMef2-gain	NA	CR43276
chr2L:21320103-21320534	nonMef2-gain	NA	crc
chr2L:21321532-21322141	nonMef2-gain	NA	crc
chr3L:15534546-15534959	nonMef2-gain	NA	CrebA
chr3L:8833340-8833642	nonMef2-gain	NA	dally
chr2L:19320685-19321628	nonMef2-gain	NA	dnt
chr2L:2830565-2831092	nonMef2-gain	NA	Duox
chr2R:6116603-6117017	nonMef2-gain	NA	EcR
chr2R:17436309-17436727	nonMef2-gain	NA	EDTP
chr3L:21256696-21257343	nonMef2-gain	NA	Eip78C
chr3L:21251246-21251937	nonMef2-gain	NA	Eip78C
chr3L:16969591-16969996	nonMef2-gain	NA	Exn
chr2R:18236430-18236940	nonMef2-gain	NA	fj
chr2R:18232579-18232935	nonMef2-gain	NA	fj

chr2L:3632025-3632487	nonMef2-gain	NA	for
chr2R:22317942-22318314	nonMef2-gain	NA	Gp150
chr3R:18989349-18989829	nonMef2-gain	NA	gukh
chr2L:21426524-21427127	nonMef2-gain	NA	His1:CG33804
chr2R:16338847-16339288	nonMef2-gain	NA	Hmgs
chr3R:10656569-10657115	nonMef2-gain	NA	hth
chr3L:3384241-3384689	nonMef2-gain	NA	Ids
chr3R:19383511-19383973	nonMef2-gain	NA	Ino80
chr3R:19382805-19383448	nonMef2-gain	NA	Ino80
chr2R:6459436-6460285	nonMef2-gain	NA	jing
chr2L:244670-244970	nonMef2-gain	FALSE	kis
chr2L:13571231-13571541	nonMef2-gain	NA	kuz
chr2L:13570097-13571089	nonMef2-gain	NA	kuz
chr3R:6671830-6672295	nonMef2-gain	FALSE	lab
chr3R:6666988-6667366	nonMef2-gain	FALSE	lab
chr2R:21733522-21733844	nonMef2-gain	NA	LBR
chr2R:14008982-14009349	nonMef2-gain	NA	mam
chr2L:121378-121811	nonMef2-gain	NA	ND-15
chr2R:17846460-17846762	nonMef2-gain	NA	Orai
chr2R:17851102-17851686	nonMef2-gain	NA	Orai
chr3R:17703905-17704328	nonMef2-gain	NA	osa
chr2R:17529928-17530302	nonMef2-gain	NA	Patronin
chr2R:23666882-23667198	nonMef2-gain	NA	Pde8
chr2L:6328903-6329203	nonMef2-gain	NA	PDZ-GEF
chr2L:3476422-3476875	nonMef2-gain	FALSE	Pgant4
chr3R:5028325-5028919	nonMef2-gain	NA	plh
chr3R:23325718-23326410	nonMef2-gain	FALSE	pnt
chr2R:8655187-8656002	nonMef2-gain	FALSE	ptc
chr3L:1407040-1407820	nonMef2-gain	NA	Ptp61F
chr3L:1407947-1408566	nonMef2-gain	NA	Ptp61F
chr3R:8854392-8854875	nonMef2-gain	NA	pyd
chr2R:9181189-9181489	nonMef2-gain	NA	Rab32
chr3R:24570574-24571046	nonMef2-gain	NA	REPTOR
chr3L:12056306-12056717	nonMef2-gain	NA	rols
chr3L:18402840-18403248	nonMef2-gain	FALSE	rpr
chr3L:18402065-18402654	nonMef2-gain	FALSE	rpr
chr3L:18403423-18403723	nonMef2-gain	FALSE	rpr
chr3L:11586681-11587129	nonMef2-gain	NA	rt
chr2L:5004752-5005101	nonMef2-gain	NA	Rtnl1
chr2L:5005127-5005634	nonMef2-gain	NA	Rtnl1
chr2R:22676703-22677302	nonMef2-gain	NA	RYBP
chr2R:18293008-18293517	nonMef2-gain	NA	sbb

chr3L:7360961-7361353	nonMef2-gain	NA	Sec63
chr3R:27193866-27194416	nonMef2-gain	NA	Ser
chr3L:14614179-14614618	nonMef2-gain	NA	shd
chr2R:14307320-14307739	nonMef2-gain	NA	Shrm
chr2R:14307921-14308350	nonMef2-gain	NA	Shrm
chr2R:19619565-19620605	nonMef2-gain	NA	sm
chr3R:28859005-28859465	nonMef2-gain	NA	spg
chr2L:19572466-19572823	nonMef2-gain	NA	spi
chr2L:3479615-3480207	nonMef2-gain	NA	Thor
chr3R:26804102-26804558	nonMef2-gain	NA	Tl
chr2L:3107952-3108652	nonMef2-gain	FALSE	toc
chr3L:15227325-15228237	nonMef2-gain	NA	Tollo
chr3L:13118970-13119558	nonMef2-gain	FALSE	trn
chr3L:13120085-13120520	nonMef2-gain	FALSE	trn
chr3L:13125319-13125723	nonMef2-gain	NA	trn
chr2L:6694054-6694619	nonMef2-gain	FALSE	Tsp
chr2R:7039928-7040337	nonMef2-gain	NA	Tsp42Ej
chr2R:7045573-7045945	nonMef2-gain	NA	Tsp42El
chr4:908926-909241	nonMef2-gain	NA	unc-13
chr3L:12111009-12111648	nonMef2-gain	NA	vers
chr2L:5299463-5299763	nonMef2-gain	FALSE	vri
chr2R:22311791-22312215	nonMef2-gain	NA	wdp
chr2R:22379375-22379740	nonMef2-gain	NA	wrapper
chr2R:20944761-20945065	nonMef2-loss	NA	Act57B
chr3L:17755689-17755989	nonMef2-loss	NA	Adgf-A
chr3R:30145878-30146299	nonMef2-loss	NA	AdoR
chr2L:2180703-2181042	nonMef2-loss	FALSE	aop
chr3R:17348702-17349032	nonMef2-loss	NA	beat-IIa
chr2R:24311093-24311393	nonMef2-loss	NA	betaTub60D
chr2L:17466058-17466450	nonMef2-loss	NA	BicD
chr2L:17466492-17467114	nonMef2-loss	NA	BicD
chr3L:9122421-9122848	nonMef2-loss	NA	bol
chr2L:8098391-8098699	nonMef2-loss	NA	Bsg
chr4:724982-725327	nonMef2-loss	NA	bt
chr2L:16167486-16167902	nonMef2-loss	NA	Ca-alpha1D
chr4:1280449-1281172	nonMef2-loss	NA	Cadps
chr4:1295851-1296264	nonMef2-loss	NA	Cadps
chr2L:20175009-20175357	nonMef2-loss	NA	CG10651
chr3L:12478249-12478753	nonMef2-loss	NA	CG10663
chr3L:13518931-13519231	nonMef2-loss	NA	CG10710
chr3L:13519564-13519961	nonMef2-loss	NA	CG10710
chr2R:20231516-20231910	nonMef2-loss	NA	CG11044

chr2L:9908429-9908863	nonMef2-loss	NA	CG13124
chr2L:16818219-16818537	nonMef2-loss	NA	CG13280
chr3L:8747300-8747759	nonMef2-loss	TRUE	CG13306
chr3L:2185641-2186052	nonMef2-loss	NA	CG13810
chr2R:19882631-19883194	nonMef2-loss	NA	CG13872
chr3L:13440794-13441370	nonMef2-loss	NA	CG14109
chr3L:20104680-20104980	nonMef2-loss	NA	CG14186
chr3R:30944356-30944785	nonMef2-loss	NA	CG15550
chr2R:9819062-9819465	nonMef2-loss	NA	CG1648
chr4:239552-239854	nonMef2-loss	NA	CG1674
chr3R:23971086-23971402	nonMef2-loss	FALSE	CG31140
chr2L:6526974-6527352	nonMef2-loss	NA	CG31637
chr3L:7718230-7718598	nonMef2-loss	NA	CG32373
chr2L:401999-402678	nonMef2-loss	NA	CG4213
chr3R:16524400-16525088	nonMef2-loss	NA	CG42342
chr3R:22124046-22124523	nonMef2-loss	NA	CG42390
chr2L:22142680-22143126	nonMef2-loss	NA	CG42748
chr3R:15122051-15122539	nonMef2-loss	TRUE	CG42788
chr3L:10745814-10746688	nonMef2-loss	NA	CG43245
chr3L:16533724-16534437	nonMef2-loss	NA	CG43373
chr2L:15643802-15644160	nonMef2-loss	NA	CG4587
chr2L:1163762-1164063	nonMef2-loss	NA	CG5080
chr3R:22488593-22489420	nonMef2-loss	NA	CG5376
chr3R:25852193-25852504	nonMef2-loss	NA	CG5886
chr3R:27247820-27248120	nonMef2-loss	NA	CG5984
chr3L:17657036-17657336	nonMef2-loss	NA	CG7484
chr3R:8300186-8300487	nonMef2-loss	NA	CG9626
chr3R:8300809-8301165	nonMef2-loss	NA	CG9626
chr2L:5971953-5972846	nonMef2-loss	NA	chic
chr2R:19232124-19232424	nonMef2-loss	NA	cora
chr2R:6869827-6870546	nonMef2-loss	NA	coro
chr3L:8865156-8865659	nonMef2-loss	NA	dally
chr3R:19317113-19317619	nonMef2-loss	TRUE	Dl
chr2R:13499445-13499745	nonMef2-loss	NA	drk
chr3R:19486845-19487159	nonMef2-loss	NA	Dys
chr2R:23743268-23743820	nonMef2-loss	NA	egl
chr3L:27897446-27898020	nonMef2-loss	NA	eIF4B
chr3L:18037077-18037534	nonMef2-loss	NA	Eip75B
chr3L:20932446-20933027	nonMef2-loss	NA	fng
chr3L:14545169-14545534	nonMef2-loss	NA	Gbs-70E
chr2R:18570755-18571409	nonMef2-loss	NA	GEFmeso
chr2R:10134630-10134971	nonMef2-loss	NA	gem

chr2L:5559064-5559397	nonMef2-loss	NA	GluRIIB
chr3R:20651862-20652425	nonMef2-loss	NA	GluRIID
chr2L:5944314-5944828	nonMef2-loss	NA	Gpdh1
chr2R:21201887-21202187	nonMef2-loss	NA	Gyg
chr2L:5427149-5427536	nonMef2-loss	NA	H15
chr3L:1501402-1502361	nonMef2-loss	NA	hfp
chr2L:21424947-21425284	nonMef2-loss	NA	His4:CG31611
chr3R:22052070-22052519	nonMef2-loss	NA	how
chr2R:18717870-18718310	nonMef2-loss	NA	Hs3st-A
chr3L:13974005-13974508	nonMef2-loss	NA	Hsc70-1
chr3L:13974522-13975134	nonMef2-loss	NA	Hsc70-1
chr2L:6557371-6557678	nonMef2-loss	NA	IFT52
chr2R:11702989-11703375	nonMef2-loss	NA	Ir48b
chr2R:12099061-12099701	nonMef2-loss	NA	jeb
chr2L:16364870-16365177	nonMef2-loss	NA	jhamt
chr3L:510921-511315	nonMef2-loss	NA	klar
chr2R:6736366-6736727	nonMef2-loss	NA	l(2)01289
chr2L:2918256-2918727	nonMef2-loss	NA	lilli
chr2R:20556613-20556913	nonMef2-loss	NA	lms
chr2R:21784318-21784897	nonMef2-loss	NA	Loxl2
chr2R:21783688-21784169	nonMef2-loss	NA	Loxl2
chr2L:6772725-6773035	nonMef2-loss	NA	LUBEL
chr2R:8462238-8462594	nonMef2-loss	NA	Mal-A5
chr2L:8080917-8081217	nonMef2-loss	NA	Mcr
chr2L:16766036-16766336	nonMef2-loss	NA	Mhc
chr3R:16091337-16091638	nonMef2-loss	NA	Mhcl
chr2L:19036270-19036827	nonMef2-loss	NA	mib2
chr2L:7042704-7043078	nonMef2-loss	NA	milt
chr2R:6961195-6961702	nonMef2-loss	TRUE	mim
chr2L:20483903-20484220	nonMef2-loss	NA	mir-1
chr3L:10353078-10353378	nonMef2-loss	NA	mir-276a
chr2R:24076099-24076460	nonMef2-loss	NA	Mlp60A
chr2R:24686205-24686568	nonMef2-loss	NA	Mmp1
chr2L:5186547-5187048	nonMef2-loss	NA	Msp300
chr2L:5100636-5100940	nonMef2-loss	NA	Msp300
chr3L:9432321-9432622	nonMef2-loss	FALSE	MTF-1
chr3R:7646355-7646766	nonMef2-loss	NA	Nlg3
chr3L:14220813-14221469	nonMef2-loss	NA	nuf
chr3L:7903382-7903909	nonMef2-loss	NA	pbl
chr2R:9861820-9862183	nonMef2-loss	NA	PCB
chr2R:9860626-9860992	nonMef2-loss	NA	PCB
chr2R:9859928-9860261	nonMef2-loss	NA	PCB

chr3R:31300424-31300842	nonMef2-loss	NA	pHCl-2
chr2L:1118072-1118483	nonMef2-loss	NA	Pino
chr2L:16761983-16762422	nonMef2-loss	NA	ppk17
chr2L:16761010-16761310	nonMef2-loss	NA	ppk17
chr2R:22925840-22926241	nonMef2-loss	NA	ReepA
chr2R:22641321-22641621	nonMef2-loss	NA	RpS24
chr2R:20940368-20940678	nonMef2-loss	NA	Rx
chr2R:8091496-8092050	nonMef2-loss	NA	sand
chr2L:6818991-6819495	nonMef2-loss	NA	sens-2
chr2L:19847591-19848140	nonMef2-loss	NA	sick
chr2L:19898856-19899158	nonMef2-loss	NA	sick
chr3R:27415047-27415648	nonMef2-loss	NA	side
chr3L:5666358-5666923	nonMef2-loss	NA	sif
chr3L:12275219-12275519	nonMef2-loss	FALSE	Sms
chr3L:4624420-4624812	nonMef2-loss	NA	Src64B
chr3R:16256072-16256523	nonMef2-loss	FALSE	tara
chr3R:8719135-8719435	nonMef2-loss	NA	TMEM216
chr3R:30545665-30545974	nonMef2-loss	NA	tmod
chr3R:30544876-30545541	nonMef2-loss	NA	tmod
chr3R:30544178-30544627	nonMef2-loss	NA	tmod
chr3L:6942148-6942514	nonMef2-loss	NA	tow
chr2R:5055397-5055867	nonMef2-loss	NA	TpnC41C
chr2R:5049625-5050427	nonMef2-loss	NA	TpnC41C
chr3L:17045656-17045956	nonMef2-loss	NA	TpnC73F
chr2R:21084033-21084393	nonMef2-loss	NA	Treh
chr3R:31718394-31718694	nonMef2-loss	NA	ttk
chr2L:18826571-18826871	nonMef2-loss	FALSE	Ugt36D1
chr2R:14243632-14243990	nonMef2-loss	TRUE	Usp20-33
chr3L:5754800-5755100	nonMef2-loss	NA	Usp47
chr2L:7267560-7268265	nonMef2-loss	NA	Wnt4
chr2R:15773041-15773412	nonMef2-loss	NA	Zasp52
chr3L:2138497-2138806	nonMef2-loss	NA	zormin

Acknowledgments

First of all, I want to express my gratitude to Eileen for giving me the opportunity to join such an amazing lab and to work on many exciting projects. Thank you for your mentorship and your continuous support throughout my PhD.

I'm thankful to my TAC members: Dr. Justin Crocker, Dr. Oliver Stegle and Dr. Karsten Rippe for their valuable advice during my PhD. Thank you Dr. Steffen Lemke for agreeing to be the Vorsitz of the defense committee.

I want to thank all the great people in the Furlong lab - both past and present members - thank you for always being friendly and making me feel welcome in the lab. Thank you for all the scientific advice and all the fun times as well!

In particular I want to thank James Reddington, who provided a lot of help and inspiration at the beginning of my PhD. You are an amazing scientist and a great friend!

A special thanks goes to our fantastic tech team: Raquel, Rebecca and Katharina - I don't think I would have made it without all your help.

I want to thank all the members of the GeneCore (Vladimir and colleagues), Flow Cytometry (Malte, Diana and Beata) and PepCore facilities and Alessandra Reversi for always providing great assistance.

Thank you, Monica, for always being by my side and reminding me of the life we have outside the lab. I'm thankful for all the great times we've had together and I can't wait for our next adventure.

Dedico questa tesi alla mia famiglia e in particolare ai miei genitori, Marco e Cristina. Grazie di aver creduto in me e di avermi sempre supportato nelle mie scelte. Grazie dei vostri insegnamenti e di tutti i sacrifici che avete fatto per permettermi di arrivare fino a qui. Senza di voi non avrei mai raggiunto questo importante traguardo.