UNIVERSITY OF CALIFORNIA,
IRVINE

Mapping transcriptional regulation of cell types and states using systems genetics in mouse

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Mathematical, Computational, and Systems Biology

by

Elisabeth Rebboah

Dissertation Committee:
Professor Ali Mortazavi, Chair
Professor Kyoko Yokomori
Professor Kim Green

2024

DEDICATION

To Mom

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

None of the work presented here could have been accomplished without my fellow lab members, past and present. First and foremost, heartfelt thanks to Fairlie, Narges, Jasmine, and Heidi. Special thanks to Fairlie for her camaraderie, coding wizardry, stellar presentations, and being an amazing co-TA. Thanks to Narges for being a wonderful project partner and brilliant bioinformatician. Thanks to Jaz, my co-TA and fellow experimentalist/bioinformatician hybrid, for her inspirational dedication and grace in everything she does, and for all the emotional support. Special thanks to Heidi, our lab manager, the true foundation of the lab, and dear friend to everyone. Her leadership, foresight, and diligence keeps us afloat during the hardest times. Thanks to Sorena and Gaby, previous members and the best bay-mates anyone could ask for, ever-helpful mentors, and exceptional scientists. Thanks to all other previous members as well, including Kate, Christina, Dana, Cassie, Camden, and in particular Bea, for her friendship and guidance over the years, and Isa, for her infectious passion for biology and mentorship. Lastly, heartfelt thanks to newer lab members, in particular Ghassan, whose remarkable aptitude and enthusiasm destines them for success. Everything we accomplish is possible because of them and the rest of the IGVF team (Parvin, Negar M., Romina, Erisa, Negar F., and Maggie). Thanks to our newest PhD students Ryan, Yeon, and Elnaz for your unwavering support. Each of you has already made remarkable contributions to the lab and it's been an honor to mentor the new generation of grad students. If I can do it you can do it!

Sincere thanks to my collaborators at UCI and Caltech. Thanks to Shimako Kawauchi, Grant MacGregor, and Brian Williams for their biological expertise, mentorship, and guidance in our IGVF project. Special recognition goes to Brian Williams and Diane Trout at Caltech, who enable excellent science through their experimental and computational skills as well as their genuine kindness. I was very lucky to have their constant support throughout ENCODE and IGVF, along with mentorship (and often a voice of reason) from Barbara Wold. I am also thankful to Lior Pachter and his lab, past and present, especially Sina Booeshaghi and Delaney Sullivan, for their invaluable help and expertise in all things single-cell.

My deepest gratitude goes to my mentor, Ali, for his guidance, unwavering support of his students, and commitment to push us to reach our full potential. Thanks to his dedication to outreach, I had the privilege of being a TA for high school students for the past three years. Their curiosity, excitement, and ambition left me with a revitalized appreciation for science every summer. Thanks also to Stephanie Shirey for her enduring friendship and support over the years.

Special thanks to my family, including our beloved dog Molly. The most important acknowledgement of all goes to my mom, as I would never have set foot in a lab without the opportunities she provided for me. Thanks to my uncle, whose love of mentoring and his commitment to his students are a constant source of motivation for me. My family deserves immense appreciation for their encouragement and support throughout the past six years. Finally, I acknowledge my grandmother, whose mathematical intuition partly inspired my pursuit of this degree.

VITA

Elisabeth Rebboah

EDUCATION

Ph.D. in Mathematical, Computational, and Systems Biology **2024**
University of California, Irvine *Irvine, CA*

B.S. in Bioengineering **2015**
University of California, San Diego *San Diego, CA*

PUBLICATIONS

* These authors contributed equally

Published

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

2. S. Pieraut, N. Gounko, R. Sando III, W. Dang, **E. Rebboah**, S. Panda, L Madisen, H. Zeng, A. Maximov. Experience-Dependent Remodeling of Basket Cell Networks in the Dentate Gyrus. *Neuron.* (2014)

In review

1. N. Rezaie, **E. Rebboah**, B.A. Williams, H.Y. Liang, F. Reese, G. Balderrama-Gutierrez, L.A. Dionne, L. Reinholdt, D. Trout, B.J. Wold, A. Mortazavi. Identification of robust cellular programs using reproducible LDA that impact sex-specific disease progression in different genotypes of a mouse model of AD. *bioRxiv.* (2024)

2. F. Reese, B.A. Williams, G. Balderrama-Gutierrez, D. Wyman, M.H. Çelik, **E. Rebboah**, N. Rezaie, D. Trout, M. Razavi-Mohseni, Y. Jiang, B. Borsari, S. Morabito, H.Y. Liang, C. McGill, S. Rahmanian, J. Sakr, S. Jiang, W. Zeng, K. Carvalho, A. Weimer, L.A. Dionne, A. McShane, K. Bedi, S. Elhajjajy, S. Upchurch, J. Jou, I. Youngworth, I. Gabdank, P. Sud, O. Jolanki, J.S. Strattan, M.S. Kagda, M.P. Snyder, B.C. Hitz, J.E. Moore, Z. Weng, D. Bennet, L. Reinholdt, M. Ljungman, M.A. Beer, M.B. Gerstein, L. Pachter, R. Guigó, B.J. Wold, A. Mortazavi. The ENCODE4 long-read RNA-seq collection reveals distinct classes of transcript structure diversity. *bioRxiv.* (2023)

3. K. Carvalho, **E. Rebboah**, C. Jansen, K. Williams, A. Dowey, C. McGill, A. Mortazavi. Uncovering the Gene Regulatory Networks Underlying Macrophage Polarization Through Comparative Analysis of Bulk and Single-Cell Data. *bioRxiv.* (2021)

4. G. Balderrama-Gutierrez, H.Y. Liang, N. Rezaie, K. Carvalho, S. Forner, D. Matheos, **E. Rebboah**, K.N. Green, A.J. Tenner, F. LaFerla, A. Mortazavi. Single-cell and nucleus RNA-seq in a mouse model of AD reveal activation of distinct glial subpopulations in the presence of plaques and tangles. *bioRxiv.* (2021)

In preparation

1. **E. Rebboah**, N. Rezaie, B.A. Williams, A.K. Weimer-Lindeboom, H.Y. Liang, F. Reese, D. Trout, J. Jou, I. Youngworth, S. Morabito, M.P. Snyder, B.J. Wold, A. Mortazavi. The ENCODE mouse postnatal developmental time course identifies regulatory programs of cell types and cell states. *In preparation.* (2024)

SELECTED PRESENTATIONS

IGVF Consortium Meeting **2023**
Accepted speaker

Biology of Genomes **2023**
Accepted speaker

ENCODE Consortium Meeting **2022**
Accepted speaker

WORKSHOPS

UCI GenPALS Intro to scRNA-seq Workshop **2022, 2023**
UCI Single cell RNA-seq Short Course **2020**

TEACHING EXPERIENCE

Teaching Assistant **2021, 2022, 2023**
UCI COSMOS (Genes, Genomes, and Skeletal Muscle Dystrophies) *Irvine, CA*

Teaching Assistant **2021, 2022**
Intro to Precision Medicine (D132) *Irvine, CA*

Teaching Assistant **2019**
UCI COSMOS (Tissue and Tumor Biology and Mathematical Modeling) *Irvine, CA*

ABSTRACT OF THE DISSERTATION

Mapping transcriptional regulation of cell types and states using systems genetics in mouse

By

Elisabeth Rebboah

Doctor of Philosophy in Mathematical, Computational, and Systems Biology

University of California, Irvine, 2024

Professor Ali Mortazavi, Chair

Complex traits are intricately intertwined with an organism's genome, a relationship underscored by the dynamic landscape of its transcriptome. Selective gene expression regulates cell type specialization and fluctuation of cell states. The development of RNA sequencing has facilitated the capture of the whole transcriptome of a given sample. However, a bulk approach obscures cell type heterogeneity, impeding the precise dissection of cell-specific effects, including those modulated by genotype, developmental stage, and disease state. In contrast, single-cell and single-nucleus RNA-seq preserves cellular identity, enabling a comprehensive mapping of gene expression across various cell types and states.

Here, I describe my work in single-cell transcriptomics to characterize cell types and cell states in mouse. First, I present our long-read single-cell RNA-seq method, benchmarked in the C2C12 mouse myogenic system, which revealed cell type-specific isoform switching in key genes during myogenesis. Next, I characterize 5 mouse tissues at single-nucleus resolution during postnatal development using the ENCODE4 mouse dataset, where I used topic modeling to reveal cell type- and state-specific cellular programs. Lastly, I investigate the impact of genetic variation on gene expression across 8 diverse tissues from 8 mouse genotypes, pinpointing genotype- driven variation in specific celltypes in both wild-derived and classical lab strains. Together, these projects lay the groundwork for cohesive cell type and

cell state annotation and comparative analyses, contributing to future characterization of these tissues in other contexts such as human diseases and hybrid mouse genotypes.

Chapter 1

Introduction

1.1 Abstract

Systems genetics integrates genetics, genomics, systems biology, and computational biology to elucidate the genetic underpinnings of complex traits and diseases. Single-cell RNA-seq has emerged as an invaluable tool for characterizing gene expression in heterogeneous tissues, allowing for the identification of distinct cell types and states. The increasing throughput of single-cell functional genomics assays has enabled large-scale studies of genomes and transcriptomes in individual cells. This enables a systems genetics approach of profiling molecular traits, such as gene expression, in specific cell types of genetically diverse populations in order to identify genomic variation associated with transcriptional regulation. Here, I discuss the use of mouse models for systems genetics within the framework of functional genomics assays. I provide a brief review of single-cell genomics technologies for measuring gene expression as well as open chromatin and offer an overview of data processing and cell type annotation methods. Finally, I explore the biological impact of single-cell and single-nucleus RNA-seq for capturing cell types and cell states in skeletal muscle, highlighting approaches

for the analysis of gene regulatory programs.

1.2 Introduction

Investigation of complex traits using systems genetics

The history of systems genetics is deeply intertwined with classical genetics and the advent of molecular biology techniques. Its roots lie in the pioneering work of researchers such as Gregor Mendel, who established the fundamental laws of inheritance through experiments with pea plants in the 19th century.[1]. Mendel's pioneering work on inheritance, published in 1865 but not widely recognized until the turn of the century, spurred a rediscovery of his principles between 1900 and 1903[1]. This revival led to a surge of significant publications across Europe, notably by plant geneticists such as Hugo de Vries, Carl Correns, and Erich von Tschermak, and further popularized by famous (and infamous) figures such as biologist William Bateson and eugenicist Charles Davenport in England and America, respectively[1]. Ronald Fisher's 1918 publication established mathematical models of genetic effects, including the infinitesimal model, proposing that quantitative traits stem from many genes with small, independent effects[2]. This concept, known as polygenicity, became a cornerstone of quantitative genetics. Theodosius Dobzhansky's 1937 experimental findings validated Fisher's theories, advancing our comprehension of genetic variation and speciation[3]. In 1953, James Watson, Francis Crick, and Rosalind Franklin elucidated the structure of DNA, providing the molecular basis for understanding genetics[4,5]. In the mid-20th century, genetic mapping and quantitative genetics techniques laid the groundwork for systems genetics[6]. A quantitative trait locus, or QTL, is a segment of DNA associated with variations in measurable characteristics, known as quantitative traits[7]. These traits encompass a range of clinical features such as height, weight, and blood pressure, as well as predisposition to diseases such as diabetes and cancer, that can be traced to molecular intermediates such as transcript expression, protein levels,

and chromatin accessibility.

One of the early milestones in systems genetics came in the late 1990s with work in fruit flies (*Drosophila melanogaster*). QTL analysis was conducted to identify genomic regions associated with variation in morphological traits such as wing shape and bristle number, providing insights into the genetic architecture of complex traits[8–10]. Examination of life span in *D. melanogaster* through QTL mapping revealed numerous alleles influencing longevity, along with significant findings of dominance, epistatic, and genotype-by-environment effects on life span[11–13]. Similarly, in yeast, pioneering studies identified QTLs associated with various phenotypes, including glucose metabolism and sporulation efficiency[14,15]. Among this research came one of the first genome-wide association studies (GWAS) in early 2005 to identify genetic variants associated with gene expression differences[16]. This study, which profiled the expression of 5,700 genes in yeast using microarray techniques, found that most detected QTLs had weak effects, highlighting the extensive genetic complexity underlying gene expression differences[16]. These findings underscore the necessity of comprehensive omics techniques to fully elucidate the intricate genetic architecture of complex traits.

More recently, the integration of omics data, including genomics, transcriptomics, proteomics, and metabolomics, has been instrumental in advancing systems genetics. The advent of NGS (next-generation sequencing) has been transformative, facilitating high-throughput profiling of genetic variants, gene expression, and molecular intermediates on a truly genome-wide scale. Collaborative initiatives such as ENCODE[17] and IGVF[18] have leveraged these advancements to comprehensively map functional elements genome-wide and quantify the impact of genomic variation on molecular traits across diverse tissues and disease contexts. This approach provides a comprehensive understanding of biological systems, elucidating the intricate interplay between genes, regulatory elements, and cellular processes. As systems biology emerged, focusing on the holistic examination of biological networks, systems genetics evolved as a fusion of genetics and systems biology. Today, driven by tech-

nological advancements and interdisciplinary collaborations, systems genetics continues to unravel the complexities of genetic systems and their role in shaping phenotypic traits. A particularly promising of application of systems genetics is to understand mammalian gene regulation using mouse strains.

A brief history of lab mice

Mice are widely recognized as excellent mammalian models for human biology due to their genetic similarity (80% of human protein-coding genes have a mouse orthologue[19]), analogous organ systems, short reproductive cycle, and genetic manipulability, facilitating disease modeling. The use of mice in scientific research traces back centuries. In eighteenth-century Japan, mice were not only kept as pets but also bred intentionally to produce desired coat and eye colors[20]. During the Edo period, Japanese breeders began performing crosses and documenting the resulting phenotypes in domesticated mice as early as 1787[20]. In Europe, the groundwork for mouse genetics research was laid by French biologist Lucien Cuénot in 1902[21]. Cuénot demonstrated that mice inherit coat colors according to Mendel's laws of inheritance, and also identified the first lethal genetic mutation in mice[21].

At the same time, William E. Castle at the Bussey Institute at Harvard published a paper on coat color genetics in mice[22]. Castle, who was the first American geneticist to use mice to study Mendelian inheritance in mammals, mentored Clarence Cook Little, another key figure in mammalian genetics[23]. Little focused on inheritance patterns, transplants, and grafts, establishing the first inbred mouse strain, DBA, in 1909. This strain harbored alleles for various coat colors, including dilute (D), brown (B), and non-agouti (A), and laid the foundation for subsequent inbred strains[23]. Following World War I, Little accepted a position at the Cold Spring Harbor Laboratory in New York, where he continued cancer research using inbred mice[23]. There, he encountered a setback when his mouse colony was decimated by a paratyphoid epidemic[24]. To rebuild his colony, Little imported albino mice maintained by Halsey Bagg at Memorial Hospital in New York City, which later became known as the BALB

strain[24]. Other strains, including C3H, CBA, and A, were developed around the same time. At Cold Spring Harbor, Litte's colleague Leonell C. Strong developed C3H from a cross of a Bagg albino female with a DBA male[24]. In 1921, he introduced the CBA strain from a cross of an unpedigreed Bagg albino female and an early DBA progenitor male, as well as the A strain from crossing a Cold Spring Harbor albino and a Bagg albino[24,25]. Simultaneously, Little established the C57BL strain, characterized by its dark fur and docile nature, sourced from a colony owned by mouse fancier Abbie Lathrop[26].

In 1929, C. C. Little officially founded the Jackson Laboratory in Bar Harbor, Maine[27]. The laboratory began as a modest summer field laboratory on Mount Desert Island. The island's climate, ideal for mouse husbandry, led to the establishment of the Roscoe B. Jackson Memorial Laboratory in memory of Little's friend and investor[27]. However, fate took a remarkable turn in October of that year. The U.S. stock market crashed, resulting in a catastrophic loss of funding for the laboratory. Despite facing severe financial constraints, scientists at the Jackson Laboratory persevered, achieving significant milestones. These include the discovery of a cancer-causing mouse mammary tumor virus and the successful execution of the first transfer of fertilized ova[27,28]. By 1937, strain 6 was isolated from the C57BL colony, giving rise to the quintessential C57BL/6J or "Black 6" strain, denoted by the /J to signify its origin at the Jackson Laboratory[29]. Today, the Jackson Laboratory (referred to as JAX) is one of the world's leading suppliers of inbred mice for biomedical research, offering various strains adapted to controlled environments to minimize genetic variability within experiments[25,27].

C57BL/6J is the most widely used strain today because of its breeding characteristics, longevity, and resistance to tumors[25]. In 2002, it became the first mouse strain to have its genome sequence published, solidifying its status as the most widely used mouse strain in biomedical research[29,30]. Sequencing has allowed insights into its genetic traits, such as a spontaneous mutation impacting glucose homeostasis[31], which may be linked to diet-

induced obesity. JAX also offers hundreds of lab strains, some with clinically relevant phenotypes, such as NOD/ShiLtJ ("Non-obese diabetic") and NZO/HlLtJ ("New Zealand obese"), used for type 1 and type 2 diabetes research, respectively[32,33]. While C57BL/6J is tumor-resistant, others are more susceptible to cancer such as the long-standing A/J (albino) strain, a well-established model for lung cancer[34], asthma[35], emphysema[36], and age-onset muscular dystrophy[37]. Others are preferred for specific lab techniques, such as stem cell derivation. Interestingly, *M. musculus* is the only known species with both permissive and non-permissive strains; while blastocysts from lab strains such as 129S1/SvImJ can be readily manipulated to form embryonic stem cells, NOD/ShiLtJ is more difficult or "non-permissive" to ESC derivation[38]. However, through careful differentiation protocols ESC lines have been established for all the strains mentioned above, including the previously non-permissive NOD/ShiLtJ strain[39]. In conclusion, the historical journey of using mice as models for scientific research reflects a remarkable convergence of scientific discovery and perseverance. From their early domestication to the establishment of heavily used strains and institutions such as the Jackson Laboratory, mice have remained indispensable allies in studying human genetics.

Genetic and phenotypic variation in diverse mouse strains

The natural genetic variation present across diverse inbred mouse strains, coupled with significant advancements in sequencing technology, has positioned them as invaluable tools for systems genetics. While human GWAS have demonstrated noteworthy success[40], a major challenge of systems genetics is the high number of genetically distinct individuals needed for mapping traits in the genome at high resolution, defined as less than 1 Mbp and down to 1 gene if possible. This task remains challenging due to the complex and polygenic nature of many traits[41]. Human genetic studies face unique challenges. The vast genetic diversity among human populations, combined with environmental influences and complex gene-environment interactions, necessitates large sample sizes to achieve statistical power and detect meaningful associations. Moreover, ethical considerations and practical limitations

often restrict the extent to which human populations can be studied. Unlike human studies, mouse models offer clear advantages, including controlled genetic backgrounds, shorter generation times, availability of large sample sizes, and the ability to manipulate genes in a controlled environment.

Since 2002, the *Mus musculus* genome assembly based on C57BL/6J has been updated over a dozen times, yielding a high-quality finished genome comparable in quality to the human genome[42]. Additionally, advancements in next-generation sequencing technology have enabled the publication of high-quality sequences for over 50 mouse strains and subspecies by the National Center for Biotechnology Information (NCBI)[43]. Beyond the classical lab strains mentioned above, wild-derived strains are also used for evolutionary biology research, including investigations into speciation and adaptation, as well as genetics research, such as mitochondrial DNA evolution[44,45] and meiotic recombination[46]. Classical lab strains are derived from the early crosses between European mice (*M. m. domesticus*) and Japanese mice (*M. m. molossinus*), resulting in genomes that are approximately 68% *M. m. domesticus* and 10% *M. m. molossinus*[47]. Another 6% is attributed to *M. m. castaneus* and 3% to *M. m. musculus*[47]. The subspecies diverged from a common ancestor approximately one million years ago and are captured today in wild-derived inbred strains: CAST/EiJ for *M. m. castaneus* and PWK/PhJ for *M. m. musculus*. WSB/J represents a purer *M. m. domesticus* background. While lab strains originate from domesticated mice, wild-derived strains were bred from individuals captured from wild mouse populations worldwide. For instance, CAST/EiJ originated from wild mice trapped in Thailand, PWK/PhJ from Prague, and WSB/J from Maryland[25]. Compared to the C57BL/6J reference genome, WSB/J, PWK/PhJ, and CAST/EiJ exhibit 6.04, 17.2, and 17.6 million single-nucleotide polymorphisms (SNPs), respectively, resulting in a diverse landscape of phenotypes[48]. Even without genomic analysis, observable differences exist; CAST/EiJ mice, for instance, have lighter brown fur and a smaller build compared to C57BL/6J, and are behaviorally more active and less docile than lab strains. Interestingly, CAST/EiJ are immune to flaviviruses due

to having the resistant allele of an oligoadenylate synthase gene, yet are highly susceptible to other viruses such as orthopoxviruses and influenza A[49–51]. They exhibit exceptional regenerative abilities in spinal cord and optic nerve neurons, making them a valuable model for exploring mechanisms of mammalian central nervous system regeneration[52]. In comparison to CAST/EiJ, PWK/PhJ are resistant to influenza A[51]. Among other observed traits, PWK/PhJ mice have also recently been shown to display sex-specific responses to diet-induced obesity[53]. In a study focusing on reproductive traits, CAST/EiJ males exhibited low sperm counts and poor motility, particularly when compared to PWK/PhJ and WSB/EiJ males, which displayed more favorable traits including high motility and normal morphology[54]. The study also highlighted a highly heritable phenotype in WSB/EiJ males, which had substantial vacuolization in seminiferous tubes of the testis compared to other strains[54].

Overall, the genetic diversity among *M. musculus* subspecies not only surpasses that found in the current human population but also exceeds the differences between modern humans and Neanderthals, whose last common ancestor dates back approximately ∼706,000 years[55]. Furthermore, the number of generations between these species surpasses that of primates by millions, contributing to the diversity generated by meiotic recombination[56]. The ultimate objective of systems genetics is to establish connections between specific DNA elements and changes in molecular intermediates, such as genes and proteins, that impact clinically relevant phenotypes. The naturally occurring genetic variation present in mouse strains has emerged as a foundational tool for establishing these connections, facilitated by advances in sequencing technology that can capture increasingly large genomic, transcriptomic, and epigenomic datasets at decreasing cost[6,57]. Therefore, researchers have collaboratively designed mouse panels to consolidate this naturally occurring genetic variation across strains and subspecies with a long history of research. Genetically diverse mouse panels such as the Collaborative Cross and Diversity Outbred panels have been introduced to model the naturally occurring variability observed in humans, with advantages such as replicability, stability, and control,

thus serving as valuable tools for the analysis of complex traits using systems genetics[58–62].

Mapping complex traits in mice

Mapping complex traits in mice has a rich history dating back several decades, with recombinant inbred (RI) strains playing a pivotal role. RI strains, resulting from repeated mating of siblings from two inbred parental strains over multiple generations, provide a valuable resource for mapping complex traits with greater precision due to their four-fold increase in recombination compared to single-generation maps. This increased recombination enhances the resolution of genetic mapping studies, enabling researchers to more accurately identify the genetic factors underlying complex traits[24]. The importance of RI strains was first realized by Donald Bailey in 1959, who recognized their potential utility for linkage analysis[24]. This led to the development of various sets of RI strains, such as the original CXB set, which Bailey brought to the Jackson Laboratory in 1967[24]. Subsequently, Benjamin Taylor further advanced the RI approach, creating standard sets of RI strains such as BXD and AKXD in 1973[24]. Despite their small size, typically consisting of 15–35 strains from a single pair of parental inbred lines, mouse RI panels offer a reproducible genetic background that facilitates the examination of gene-environment interactions and the use of multiple phenotyping techniques[24]. While RI strains require considerable time and effort for their development, they remain a valuable resource for investigating Mendelian and quantitative traits. Introduced in 2002, advanced mapping strategies leveraging RI intercross (RIX) strains extends the power of RI lines by producing F1 hybrids between parental RI lines, offering twice the number of recombination sites in a single individual[63].

In 2004, the Complex Trait Consortium proposed the Collaborative Cross (CC) panel, consisting of approximately 100 highly diverse RI strains originating from the eight strains mentioned in the previous section: C57BL/6J, NOD/ShiLtJ, NZO/HlLtJ, A/J, 129S1/SvImJ, WSB/J, PWK/PhJ, and CAST/EiJ (Fig. 1.1)[58]. CC lines are generated through the intermating of these parental inbred strains, followed by successive sibling mating in a "funnel"

pattern (Fig. 1.2)[64]. This process yields genetically stable lines, each representing a complex mosaic of the original genomes. Using multiple CC lines aids in statistical power to discern between true QTLs and false positives. For example, the CC lines have been used to map gene expression QTLs (eQTLs) and chromatin accessibility QTLs (caQTLs) in diverse tissues, including lung, kidney, and liver[65]. The long-term stability of established CCs as inbred strains ensures reproducible genetic backgrounds, contributing to the consistency of QTL analyses across experiments and laboratories.

Although the CCs have high statistical power for detecting genetic loci, they have limited mapping resolution, spanning tens of megabase pairs[66]. The Diversity Outbred (DO) collection, introduced in 2012, is a randomized breeding colony based on the same eight founder strains that can be leveraged for more high-precision QTL mapping[66]. For example, a 2 Mb region on chromosome 3 was mapped to serum cholesterol levels[62]. More recently, an integrative analysis of skeletal phenotypes in the the DO panel identified an eQTL impacing expression of the gene *Qsox1*, which in turn has an effect on cortical bone morphology[67]. A follow-up study mapped gene expression in bone marrow–derived stromal cells at the single-cell level, demonstrating the impact of genetic variation on the proportions of osteogenic cell types in DO mice[68]. As single-cell sequencing technologies rapidly evolve, they are expected to continue playing a vital role in the ongoing development and application of systems genetics.

The emergence of sequencing assays for functional genomics

Mapping the genetic influence on molecular traits such as transcript and protein abundance is crucial for understanding complex phenotypic traits. For example, mutations in the dystrophin gene can lead to a deficiency in the corresponding protein, a key factor in Duchenne muscular dystrophy[69]. However, many complex traits including common diseases are polygenic, meaning that each trait can be impacted quantitatively by multiple genes[41]. Rather than directly assessing protein expression, researchers often opt to measure gene expression,

or the abundance of polyadenylated RNA transcripts that serve as precursors for proteins or functional non-coding RNAs. This molecular readout provides a snapshot of ongoing cellular processes within cells. Additionally, chromatin accessibility, which indicates regions of DNA regions that are loosely bound by histones or deplete altogether, provides insights into active regulatory regions within the genome that are often associated with gene expression, enhancer activity, and transcription factor binding[70,71]. Before the introduction of illumina next-generation sequencing (NGS) in 2007[72], these assays were restricted to techniques such as microarrays, which employ a set of oligonucleotide probes to capture target molecules, relying on a predefined panel of known target genes[73]. In contrast, Illumina sequencing is a massively parallel sequencing technology that synthesizes DNA sequences base-by-base[74] allowing for comprehensive and unbiased capture of the entire genome and transcriptome.

With standard Illumina short-read sequencing, short fragments of DNA (typically 200-500 base pairs) undergo treatment to attach specific adapters at both ends[75]. These adapters are complementary to sequences anchored to a flow cell, with one end binding to the flow cell and the other remaining untethered. During a specific step of sequencing called cluster generation, these fragments are amplified while still bound to the flow cell, creating dense regions of identical sequences[75]. Sequencing proceeds as fluorescently tagged nucleotides attach base-by-base to the template, with a picture taken at each addition. Since all strands in a cluster are identical, as the same base binds to each position, a bright spot of light is generated that is detectable by a high-resolution camera. Raw image data are then translated into nucleotide sequences based on the sequence of colors (or the absence of color) at each cluster[74,75].

RNA sequencing (RNA-seq) involves extracting RNA from a cell or tissue, using an oligo-dT primer to select messenger RNAs and long non-coding RNAs[76]. This primer serves as an anchor for reverse transcription of the single-stranded RNA from the 3' tail to the 5' end, creating an RNA-cDNA hybrid[77]. As the reverse transcriptase enzyme reaches the end of

the RNA template, it adds a few non-templated nucleotides, typically cytosines[77]. Another primer, known as a template switching oligo, anneals to these non-templated nucleotides at the 5' end of the cDNA, providing the 5' primer site for subsequent PCR amplification[77]. This ensures that the amplified reverse-transcribed RNA, commonly referred to as complementary DNA or cDNA, contains the entire full-length transcript as a double-stranded, stable structure. Additional ligation, PCR, and clean-up reactions are performed to add adapters to each end of the cDNA fragment for sequencing. Importantly, for short-read Illumina sequencing, the full-length cDNA must be fragmented further in order to be readable by the sequencer, while long-read sequencers such as Oxford Nanopore can sequence full-length RNA and up to millions of bases of genomic DNA[78].

Chromatin accessibility assays also rely on sequencing DNA fragments excised from the genome. The first iteration of this assay, DNase-seq, utilizes DNase I endonuclease digestion to cleave open regions of DNA[79]. Subsequent steps involve amplification and addition of sequencing adapters to create the final library, or collection of DNA molecules that are compatible with a sequencing platform. Its successor, ATAC-seq (Assay for Transposase-Accessible Chromatin), employs a hyperactive Tn5 transposase to insert sequencing adapters into open chromatin, bypassing a library preparation step and streamlining the process[80].

The availability of multiple sequencing platforms, including Illumina, Pacific Biosciences, and Oxford Nanopore coupled with declining sequencing costs, has facilitated the proliferation of sequencing assays such as RNA-seq and ATAC-seq[57]. Consequently, the availability of publicly accessible sequencing data has surged, exemplified by initiatives such as ENCODE. Established in 2004, the goal of the ENCODE (ENCyclopedia Of DNA Elements) consortium was to build a catalog of functional DNA elements in various human and mouse cell lines, tissues, and developmental time points[17,81]. Various sequencing assays, including RNA-seq, DNase-seq, and ATAC-seq, were conducted on unified sets of samples in labs worldwide to identify regulatory elements in the genome and their associated target genes. Notable

findings from the ENCODE project include the discovery that the majority of SNPs identified by GWAS are located within DNase I hypersensitive regions, indicating the active state of crucial variants[82]. Moreover, extensive sampling uncovered the tissue-specific nature of most regulatory elements, where genomic regions accessible in one tissue often remain inaccessible in others[83]. Additionally, polyA gene expression profiling using RNA-seq during prenatal mouse development revealed distinct gene expression clusters and sub-clusters associated with specific cell types within diverse tissues[84]. In its final phase that finished in 2022, the ENCODE consortium added substantial single-cell and single-nucleus sequencing data to comprehensively capture cell type-specific signatures in human and mouse postnatal tissues.

Functional genomics at single-cell resolution

While bulk assays capture an average signature across an entire sample, single-cell and single-nucleus assays offer the capability to resolve molecular profiles, such as gene expression and chromatin accessibility, while preserving the identity of each individual cell. When scaled up to profile thousands or even hundreds of thousands of cells simultaneously, researchers can readily discern heterogeneous cell types, or groups of cells with distinct molecular profiles, and cell states, or different phenotypes within those cell types. Bulk assays can somewhat capture the dominant profile attributed to the most abundant cell type, but overlook the contributions of minor cell types, which can play crucial roles in specific developmental or disease contexts. For instance, activation of the brain's resident immune cells (microglia) is a major hallmark of disease progression in Alzheimer's disease[85]. The key to profiling many cells in a single experiment lies in ensuring that each cell is assigned a unique barcode, which is attached to all the molecules associated with that specific cell. Thanks to advances in barcoding technologies, single-cell assays are rapidly replacing bulk assays as the preferred method for large-scale genomic and transcriptomic studies wherever appropriate.

The initial phase of single-cell assays involved pipetting one cell per well in a 96-well or slightly larger format, followed by parallel experimental reactions in plate format[86]. How-

ever, these methods were both time-consuming and low-throughput. Over the past decade, advancements in droplet microfluidics have dramatically increased the number of cells per experiment[86]. For instance, the commercial Chromium system from 10x Genomics enables high-throughput profiling of RNA 3' ends from thousands of single cells; up to 80,000 in a standard chip and 320,000 in the high-throughput version[87]. This droplet-based approach uses microfluidics to encapsulate cells and barcoded beads within nanodroplets. Each synthetic bead is coated with oligonucleotide sequences containing a unique barcode. A continuous flow of cells and barcoded beads in a water-based buffer is merged with another channel containing oil, resulting in the formation of robust nanodroplets, ideally containing a single cell and one barcoded bead[86]. Subsequent cell lysis and reactions, such as reverse transcription (for RNA-seq) or Tn5 tagmentation (for ATAC-seq), occur within each droplet, releasing the barcodes from the bead and attaching them to various molecules inside the cell, such as polyadenylated RNA or tagmented DNA, depending on the capture method. However, limitations of droplet-based barcoding include access to microfluidics equipment, compatibility of cell morphology with small microfluidics channels, and cost, especially regarding large-scale experiments with many unique samples.

Ever higher throughput can be achieved using an alternative single-cell RNA-seq (scRNA-seq) method where cells serve as their own fixed containers as they undergo rounds of barcoding reactions[88]. The initial barcode, introduced during *in situ* reverse transcription, is uniquely distributed across wells in a 96-well plate, allowing for up to 96 multiplexed samples. Subsequently, the next two sets of 96 barcodes are ligated in situ following the first barcode. The cells are then counted and distributed into subpools for the final barcoding ligation, resulting in a throughput of 1 million cells with approximately 14 million possible barcodes[89]. Compared to droplet-based scRNA-seq barcoding, this method scales rapidly and requires no specialized equipment. In recent years, commercial combinatorial barcoding kits from Parse Biosciences have made scRNA-seq more accessible to many labs by offering kits in a variety of sizes (up to 1 million cells in a single experiment) and capture kits for se-

lectively targeting certain transcripts, such as immune cell type markers and exon-containing fragments[89].

While advancements in barcoding technologies are ongoing, the protocols for sample preparation present ongoing challenges. Extracting intact cells from complex tissues such as the brain is often limited to antibody-based selection techniques targeting specific cell surface markers for positive or negative selection. In many cases, researchers turn to single-nucleus sequencing as an alternative to single-cell sequencing. While this approach results in the loss of cytoplasmic transcripts, single-nucleus RNA-seq (snRNA-seq) provides a snapshot of ongoing cellular processes, enabling the identification of cell types and states as effectively as scRNA-seq[90]. For assays such as ATAC-seq that focus on genomic DNA, nuclear preparation is already integrated into the protocol. The success of any single-cell or single-nucleus assay relies heavily on achieving a clean suspension of cells or nuclei at a known concentration. This can be challenging depending on the tissue source, necessitating careful preservation of nuclear structure during tissue lysis and homogenization while minimizing non-nuclear debris. Techniques such as physical filtering through mesh strainers and density gradients aid in separating nuclei from debris based on weight. In droplet-based barcoding, input concentration is critical to avoid overloading the microfluidics equipment, which can lead to an increase in doublets (two cells with identical barcodes) and empty droplets. While cell concentration is less crucial for combinatorial barcoding, overloading can still result in an elevated risk of doublets. Although automated workflows for transitioning from whole tissue to single-cell suspension are emerging, it's essential to maintain careful bench practices to ensure high-quality downstream data.

Finally, substantial progress has been made in multiomic single-cell assays that capture two or more types of molecular profiles from individual cells simultaneously. The most widely used approach combines transcriptome profiling (RNA-seq) with chromatin accessibility (ATAC-seq), dominated by a popular 10x Multiome kit[87]. However, non-commercial methods such

as SHARE-seq, a combinatorial barcoding-based multiome assay, are also emerging[91]. In SHARE-seq, fixed cells and nuclei undergo in situ reactions where open chromatin is first excised from the genomic DNA via a Tn5 reaction, followed by reverse transcription of polyadenylated RNA[91]. Cells are barcoded through three ligation-based split-pool reactions in 96-well plates before being divided into aliquots and lysed for library preparation[91]. Other multiomic assays combine epigenetic and transcriptomic assays to capture the transcriptome and methylome, transcriptome and proteome, and even transcriptome, methylome, and chromatin accessibility simultaneously[92]. With the increasing accessibility of single-cell sequencing, there is anticipation of a continuing surge in innovative multiomic assays harnessing the power of both single-cell barcoding and high-throughput sequencing.

Single cell RNA-seq data analysis and cell type annotation

Single-cell RNA-seq libraries are typically sequenced as paired-end reads. One read contains the cell barcode and unique molecular identifier (UMI), a short sequence of random nucleotides added before PCR amplification to track unique molecules and remove PCR duplicates during data processing[93]. The other read contains the cDNA sequence. Sequencing depth, or the number of reads per cell, is crucial to fully capture the transcriptome. Sequencing saturation, a measure of the ratio of unique molecules detected to the total number of reads, helps determine whether further sequencing would provide new UMIs and additional information or merely sequence the same molecules repeatedly, thereby wasting sequencing costs.

After sequencing, the initial step in data processing involves mapping cDNA fragments to a reference, assigning reads to genes, demultiplexing cell barcodes, and deduplicating UMIs. This process yields a cell-by-gene count matrix, containing the counts of RNA molecules in each cell for each gene. Several bioinformatic tools are available for these tasks. For instance, 10x's CellRanger uses STAR for read mapping and proprietary algorithms for cell demultiplexing[87]. Recently, tools such as STARSolo[94] and kallisto bustoolskb have emerged

that enhance computational efficiency by integrating alignment, cell demultiplexing, and UMI deduplication into a unified workflow. Cell demultiplexing involves identifying cell barcodes in reads, correcting barcodes based on the known sequences used to design the oligonucleotide barcodes used in the experiment and removing duplicates. UMI deduplication compares UMIs associated with each read to eliminate duplicates originating from the same molecule, mitigating PCR amplification biases. STARSolo, based on the STAR aligner, offers high accuracy in gene quantification but requires longer processing time[94]. In contrast, kallisto bustools pseudoaligns reads to a reference transcriptome, significantly reducing processing time[95]. Both tools incorporate barcode error correction and UMI deduplication, generating a final count matrix as the final step of their processing pipeline. Importantly, intronic reads must be counted along with exonic reads when performing single-nucleus rather than whole-cell sequencing, since many transcripts are anticipated to be pre-spliced or in the process of splicing at the time of capture[96].

After generating a counts matrix, cells undergo various quality control (QC) checks such as assessing the number of UMIs per cell and the number of genes captured with at least one UMI. Additional metrics include the percentage of reads associated with mitochondrial and ribosomal transcripts. Another critical metric is the doublet score, which is an algorithm-based prediction indicating whether a cell is a doublet. This score is derived by randomly sampling and combining observed transcriptomes from single cells[97]. The local density of simulated doublets as measured by a nearest neighbor graph is utilized to calculate a doublet score for each cell[97]. Steps to calculate these QC metrics are often integrated into standard data processing workflows and are part of popular single-cell RNA-seq toolkits such as Seurat[98] and Scanpy[99], which are based on R and Python programming languages respectively. These toolkits encompass filtering cells based on QC metrics, normalizing count matrices to account for sequencing depth variation across cells, optional batch correction, principal component analysis (PCA) for linear dimensionality reduction, computing K-nearest neighbors for each cell, and clustering cells using algorithms such as Leiden and Louvain[100]. The result-

ing clusters, or groups of single cells indicate cells whose transcriptomes are similar to each other and thus suggesting a common cell type or cell state. Optional steps also involve embedding the neighborhood graph in a UMAP or t-SNE, which are nonlinear low-dimensionality representations of the data where each cell is depicted as a point in 2D space[101]. However, in recent years, the structures of UMAP embeddings have been incorrectly interpreted as biologically meaningful[102]. It is important to note that different selections of highly variable genes and other parameters, such as the metric used to construct neighborhood graphs and the number of neighbors considered for nonlinear dimensionality reduction significantly influence the resulting embedding[102]. These variations can lead to inconsistencies and misinterpretations, potentially resulting in the misassignment of cell types[102].

Cell type annotation, which is the process of assigning specific cell type labels to individual cells or clusters based on their gene expression profiles, remains challenging in single-cell RNA-seq analysis. Various bioinformatic tools have been developed to facilitate this task, including data integration methods, classification-based machine learning approaches, and marker gene databases for cell type prediction[103]. Integration techniques transfer labels from reference data to query cells within clusters but suffer from high computational demands and potential batch effects if the reference and query data vary significantly in quality or depth[104]. In contrast, supervised annotation involves constructing a classifier using a labeled reference dataset, selecting features, training the classifier, and predicting cell types in unannotated data[103]. However, reliance on reference data poses challenges as new single-cell studies grow in scale and encompass diverse biological contexts and genetic backgrounds. Manual annotation by domain experts remains popular, with clustering algorithm choice and resolution influencing the granularity of cell types. Annotators often perform differential expression analysis to identify cluster marker genes, which can indicate specific cell types. However, consideration of biological context is vital. For example, the *Prox1* homeobox transcription factor is a marker of both endothelial cells in lymphatic vasculature throughout the body[105] but is also selectively expressed in the hippocampal dentate gyrus during granule

cell (neuron) maturation[106]. Thus, annotations done in a context-aware manner with pre-set expectations of the recovered cell types often outperforms automatic annotations.

Transcriptional regulation of cell types and cell states in skeletal muscle

An organism begins as a single-celled zygote from which diverse cell types self-organize into tissues and organs during embryonic development. Organs undergo significant postnatal changes in cell type dynamics as tissues grow and mature. Notably, sex-specific cell types emerge in organs such as the adrenal gland during puberty and decline in adulthood[107]. Additionally, stem cells such as oligodendrocyte precursor cells in the brain persist throughout postnatal development to adulthood, overseeing tissue maintenance and repair[108]. Cell types are discerned by consistent, heritable features, including molecular markers, morphology, tissue location, and functional characteristics[109,110].

In skeletal muscle, a subset of *Pax7*-expressing stem cells, also known as satellite cells, rest beneath the basal lamina of myofibers and are activated to carry out myogenesis throughout the life of the organism[111]. Activated satellite cells proliferate, align, and fuse into developing or regenerating myotubes, expressing myogenic regulatory factors such as *Myf5*, *Myod1*, and *Myog*[111]. These transcription factors (TFs) control cellular fate by binding to target gene promoters such as muscle-specific actins and myosins that are essential for proper myofiber functions. Single-cell and single-nucleus RNA-seq of cultured muscle cells has allowed researchers to characterize progenitor and differentiated cell types based on their transcriptomes, revealing the temporal dynamics of gene expression during cellular differentiation[90]. In tissue, mature myonuclei are positioned along the periphery of myofibers to accommodate linearly arranged sarcomeres and interspersed mitochondria, while supporting cell types occupy the space outside of the fibers (Fig. 1.3). Single-cell RNA-seq captures non-muscle cell types such as endothelial and lymphatic endothelial cells of the blood and lymph vessels, circulating and tissue-resident immune cells, and a stromal cell type called fibro-adipogenic progenitor cells (FAPs)[112,113]. FAPs help maintain a favorable microenvironment during

muscle regeneration and repair, interacting with satellite cells and maintaining homeostasis[114]. Single-cell experiments allow for comparisons of cell type proportions, cell signaling analysis, and differential expression between cell types in various contexts such as developmental timepoint, disease, species, and sex. Alongside single-cell RNA-seq, single-nucleus ATAC-seq assesses the open chromatin landscape in single cells, determining accessibility at gene promoters and non-coding regulatory regions, or DNA sequences that can increase or decrease gene transcription[115]. With appropriate read depth, it can also be used for transcription factor footprinting to predict the binding locations of TFs, and subsequent motif enrichment analysis to match sequences in these binding locations to preferred sequence motifs for specific TFs[71]. For example, snATAC-seq footprinting in regenerating muscle revealed *Pax7* and *Nr3c1* are more active in quiescent satellite cells compared to activated satellite cells, with target genes whose expression was further validated with bulk RNA-seq[116].

Skeletal muscle is unique in that it is multinucleated, with up to hundreds of nuclei dispersed throughout the myofiber cytoplasm, presumably in part due to the substantial metabolic and functional demands required for muscle function[117]. Given this characteristic, most single-cell studies in skeletal muscle prefer to focus on individual nuclei[118]. Single-nucleus RNA-seq in skeletal muscle characterized gene expression in specialized myonuclei located beneath the neuromuscular junction (NMJ), which exhibit distinct transcriptional profiles for the development of acetylcholine receptor (AChR) clusters, essential for voluntary movement[113]. Another study combining snRNA-seq with validation using spatial data showed that all the nuclei in a muscle fiber are coordinated in their transcriptional output[119], preferring to express either slow-twitch or fast-twitch genes. Importantly, myonuclear specialization and fiber type switching are changes in cell state rather than cell type. *In vitro* experiments have demonstrated that AChR clustering can be induced in muscle cells by administration of synaptic proteins[120], while fiber type can be influenced by exercise training[121]. While cell types establish the foundation for cellular identity, cell states enable plasticity within specific cell types, exemplified by specialized NMJ and slow/fast-twitch myonuclei[110].

Modeling gene regulation in single cells

An ongoing challenge in single-cell RNA-seq analysis is identifying and associating groups of genes with meaningful traits. While gene expression differences in traits such as sex and age can be assessed through differential expression analysis, which highlights genes enriched in one group or another, such analyses offer limited insights into the relationships between different genes. Therefore, adopting a comprehensive approach is crucial for achieving a systems-level understanding of gene expression regulation. Weighted gene co-expression network analysis (WGCNA) is a commonly used approach for capturing underlying structure in data[122], reflecting the complex organization inherent in biological systems and recently adapted for single cell data[123]. However, WGCNA typically assigns each gene to a single module, limiting its flexibility. Another promising method is Latent Dirichlet Allocation (LDA), also known as topic modeling. Initially developed for population genetics and later adapted for natural language processing using machine learning, LDA groups words into topics in written documents, enabling multiple topics to be associated with a single document while assigning a numeric weight to each word in every topic[124,125]. In the context of single-cell RNA-seq, LDA treats genes as words, cells as documents, and latent biological processes as topics[126–128]. This approach aligns with true biological systems, where a gene may participate in multiple regulatory programs. By analyzing gene weights between topics, LDA facilitates the comparison of more latent traits associated with topics, such as dynamic cell types and states.

1.3 Conclusions

Single-cell RNA-seq has emerged as a transformative tool in genomics, providing invaluable insights into gene expression at the individual cell level and offering high-resolution views of heterogeneous tissues. As these assays become standard practice, they enable the study

of intricate transcriptional dynamics in specific cell types during development and assessing the impact of genetic variation on gene expression, further advancing our understanding of complex biological systems.

In Chapter 2, I describe our method for long-read single-cell RNA-seq and the benchmarking results conducted on a differentiating skeletal muscle cell line. Our results, published in *Genome Biology*, characterize heterogeneous mononucleated populations in differentiated cultures and validate cluster marker genes by RNA fluorescent *in situ* hybridization. The study highlights the gene troponin T2 (*Tnnt2*), which switches isoform expression from a short version to a long version of the transcript during differentiation, observed in both mononucleated cells and fused myotubes compared to undifferentiated myoblasts. Paired single-nucleus chromatin accessibility (snATAC-seq) data further confirms TSS switching in *Tnnt2* and several other genes, shedding light on the genomic basis of RNA isoform expression during skeletal muscle differentiation.

In Chapter 3, I provide a comprehensive overview of postnatal development across various tissues using the ENCODE4 mouse single-cell dataset, building upon previous prenatal single-cell analyses[84]. This coordinated single-nucleus RNA-seq dataset encompasses adrenal gland, left cerebral cortex, hippocampus, heart, and skeletal muscle tissues at seven postnatal timepoints. Additionally, 10x Multiome (snRNA-seq and snATAC-seq) data at two key timepoints supplements our analysis. We use LDA to model regulatory programs as topics, capturing cell type and cell state-specific signatures. The postnatal time course reveals dynamic changes, such as glial maturation in the brain and gradual expansion of skeletal muscle relative to supporting cell types such as fibro-adipogenic progenitors. Finally, we conduct an integrative analysis using the multiome data to characterize sexually dimorphic transcription factor expression and binding activity in a specific cortical layer of the adrenal gland that emerges during puberty.

In Chapter 4, I present the first phase of our project within the IGVF (Impact of Ge-

nomic Variation on Function) consortium[18] to characterize the transcriptional landscape at single-cell resolution across eight core tissues in the eight CC and DO founder genotypes (C57BL/6J, NOD/ShiLtJ, NZO/HlLtJ, A/J, 129S1/SvImJ, WSB/J, PWK/PhJ, and CAST/EiJ), with four male and four female replicates per genotype, thus christened the "8-cubed" single-nucleus RNA-seq dataset. I investigate genotype- and sex-specific clustering of single cells and explore differential gene expression between genotypes across over a hundred cell types from the same tissues profiled in ENCODE4 as well as kidney, liver, male and female gonads, and the diencephalon and pituitary brain regions. Notably, we observe early activation of satellite cells in A/J mice compared to other genotypes, which are predisposed to late onset muscular dystrophy due to a mutation in the dysferlin gene[37,129].

In Chapter 5, I outline future trajectories for the research outlined in this thesis, emphasizing the forthcoming phases of our IGVF project. I delve into the prospects of characterizing cis and trans regulation using F1 crosses, elaborate on eQTL analysis leveraging the Collaborative Cross, and explain how integrating single-cell chromatin accessibility or multiome data will bolster our QTL discovery. Additionally, I offer broader insights and aspirations for the field gleaned from navigating extensive and biologically diverse single-cell datasets.

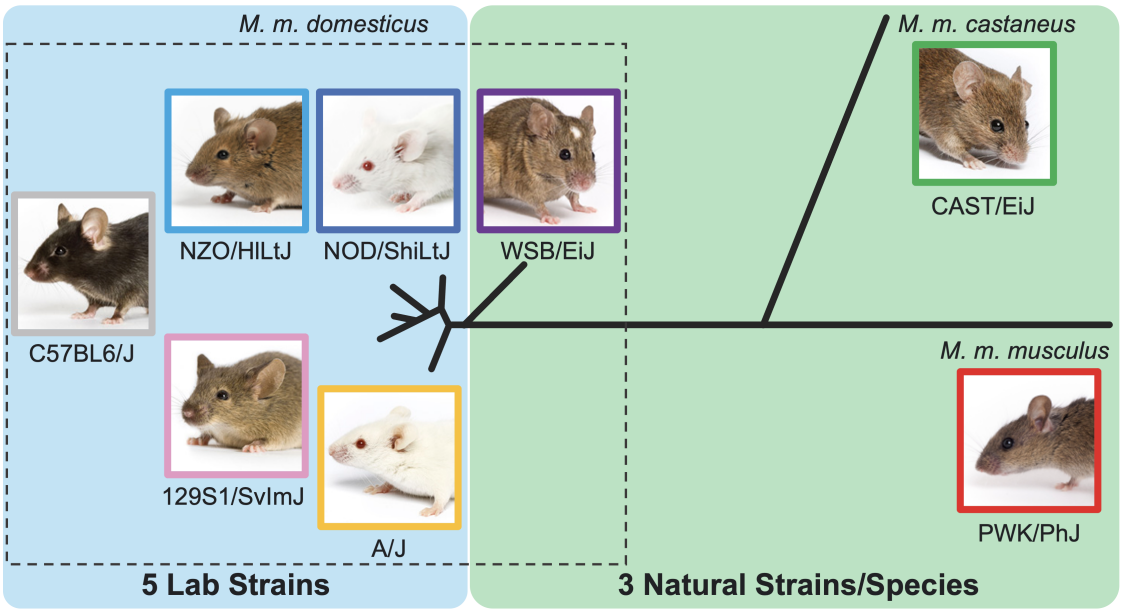

Figure 1.1: **Phylogenetic tree of the eight CC founders.** Based on single nucleotide polymorphism (SNP) data for chromosome 11[61]. Blue background indicates inbred lab strains and green background indicates wild-derived strains. Of the 3 wild-derived strains, WSB/EiJ also originates from *M. m domesticus*, while CAST/EiJ and PWK/PhJ are distinct subspecies. Mouse pictures courtesy of Jackson Laboratory.

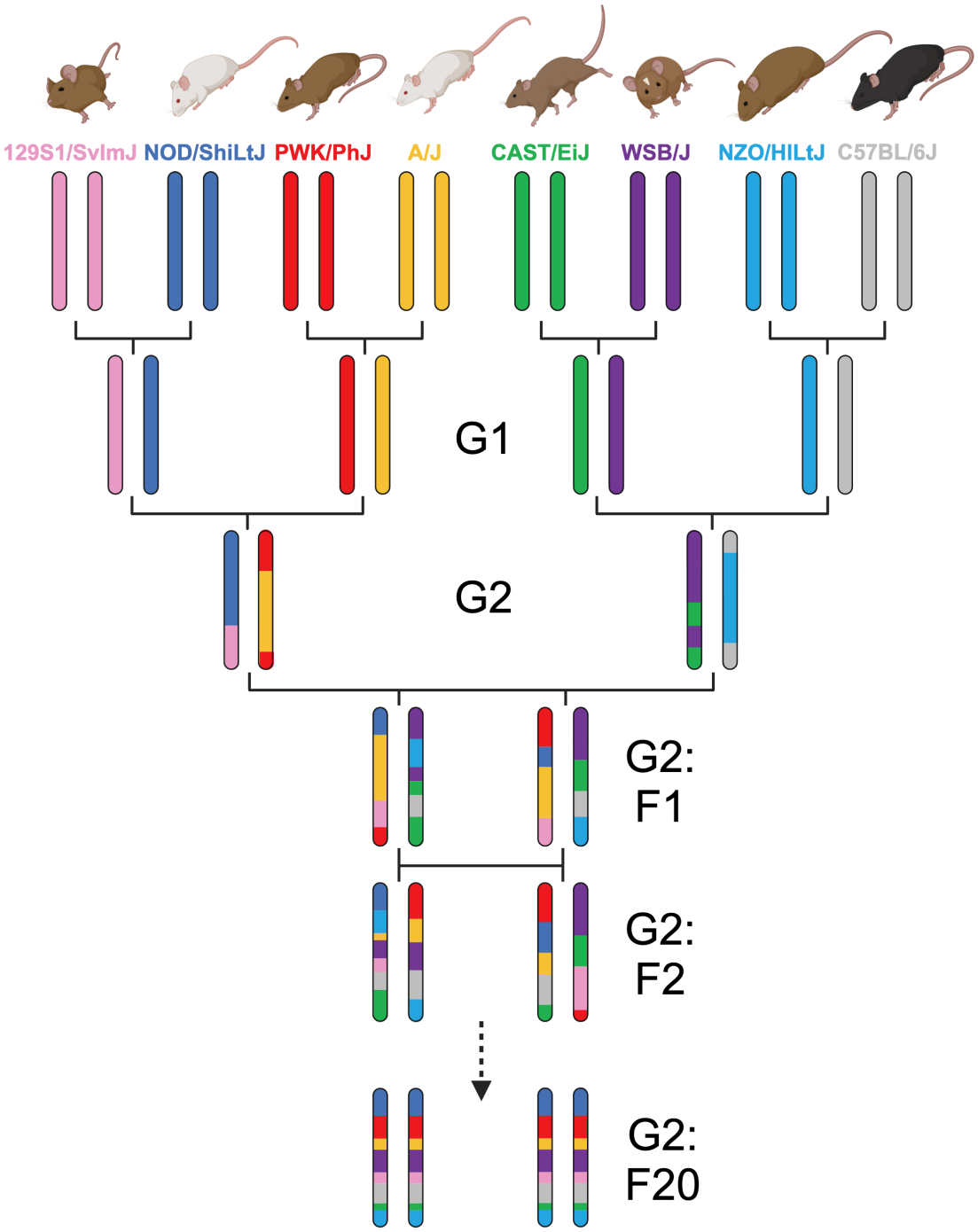

Figure 1.2: **Breeding scheme for generating a CC line.** The order of the founders is randomized and not repeated across CC lines. Initially, two of the eight founders are crossed in the outbreeding generation 1 (G1). Subsequently, those lines are intercrossed with each other in G2. Repeated generations of inbreeding through sibling mating produce filial generations until reaching (near) homozygosity[64].

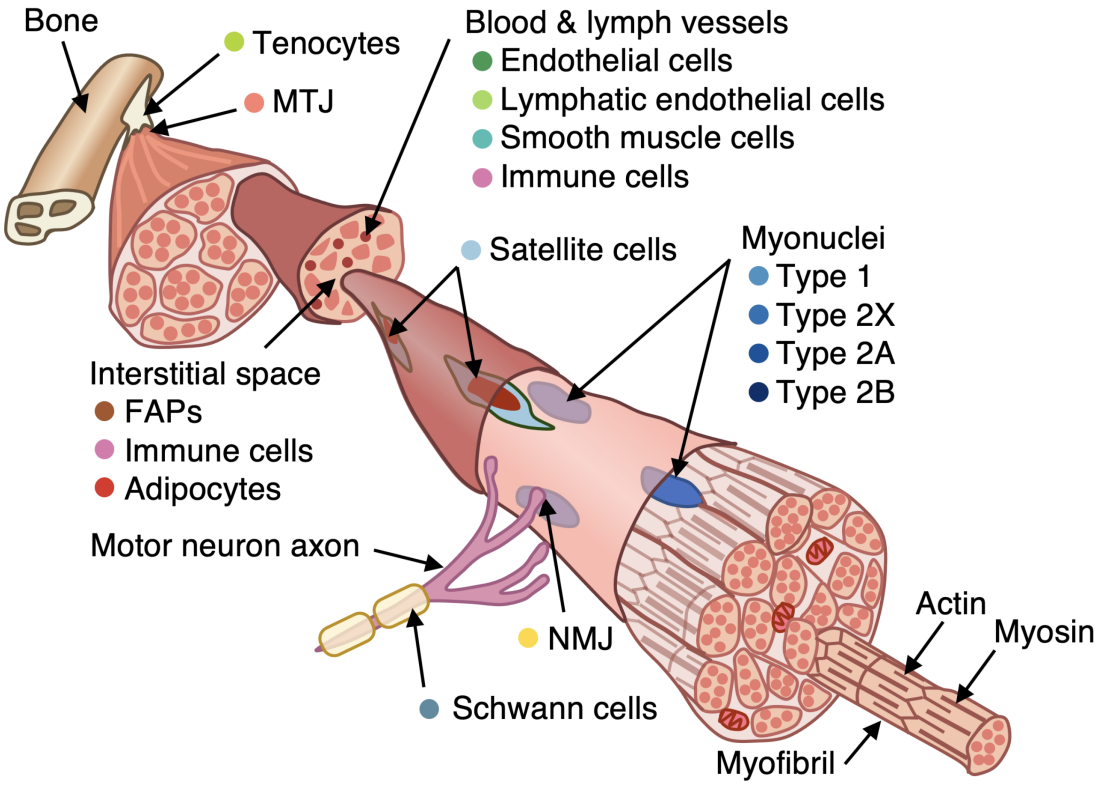

Figure 1.3: **Cell types and cell states of skeletal muscle.** A single muscle cell or myofiber contains hundreds of myonuclei, the most abundant cell type in skeletal muscle. Type 2A/B/X are subtypes of fast twitch fibers, while Type 1 are slow twitch fibers. Specialized myonuclei rest underneath the neuromuscular junction (NMJ) and myotendinous junction (MTJ), and satellite cells or muscle stem cells rest on top of the myofiber.

Chapter 2

Mapping and modeling the genomic basis of differential RNA isoform expression at single-cell resolution with LR-Split-seq

2.1 Abstract

The rise in throughput and quality of long-read sequencing should allow unambiguous identification of full-length transcript isoforms. However, its application to single-cell RNA-seq has been limited by throughput and expense. Here we develop and characterize long-read Split-seq (LR-Split-seq), which uses combinatorial barcoding to sequence single cells with long reads. Applied to the C2C12 myogenic system, LR-split-seq associates isoforms to cell types with relative economy and design flexibility. We find widespread evidence of changing isoform expression during differentiation including alternative transcription start sites

(TSS) and/or alternative internal exon usage. LR-Split-seq provides an affordable method for identifying cluster-specific isoforms in single cells.

2.2 Introduction

Alternative transcript isoform expression is a major regulatory process in eukaryotes that includes differential TSS (transcription start site) selection, RNA splicing, and TES (transcription end site) selection. These differential choices sculpt the transcriptome and its resulting proteome during development, across cell types and in disease states. However, it has proved challenging to fully capture and quantify isoform regulation by standard short-read RNA-seq because of the ambiguity it leaves in mapping the transcript termini and full-length exon connectivity that define each mature isoform.

In recent years, long-read RNA sequencing technologies have emerged as a powerful alternative for transcript-level identification and quantification by going beyond the level of exon-usage to simultaneously identify novel isoforms with alternative TSSs, TESs, and exon combinations. Furthermore, long-read RNA-seq has been adapted to single-cell sequencing using high-throughput microfluidics-based methods[130–133]. Some of these studies sequenced the same cells with both PacBio and Illumina technologies and relied on short-read gene quantification to cluster and characterize cell types, while using the long reads to identify full-length isoforms[131,133]. However, these prior approaches used expensive equipment, such as microfluidics platforms, and/or applied very high amounts of long-read sequencing whose expense limits routine and extensive application.

Differential RNA isoforms discriminate cell types within complex tissues and, within cell types such as neurons, can further distinguish functionally distinct cell sub-populations[134,135]. Isoform choice can even distinguish individual neurons of the same "type" from each other[136,137].

Transcript isoforms also discriminate developmental stages and disease states[138]. In vertebrate systems, differential isoform regulation through development has long been appreciated, and in some disease states such as type 1 myotonic dystrophy, fetal or neonatal stage isoforms of *Tnnt2*, *Atp2a1* (*Serca1*), and *Ldb3* (*Zasp*) are inappropriately expressed[139–141]. In addition, several studies have characterized the diversity of gene expression within the population of nuclei from myotubes[112,119]. This prior work on skeletal muscle provides known instances of isoform choices that we can use to benchmark new methods for transcriptome profiling, while at the same time posing unanswered questions that require single- cell or single-nucleus long-read data such as nuclear specialization within myotubes.

In vitro differentiation of the myogenic C2C12 cell line from proliferating, mononucleated myoblasts to multinucleated myotubes is a widely used model of myogenesis due to transcriptional and morphological similarities to the in vivo process[142]. A subset of cells under differentiation promoting conditions remain mononucleated and are called MNCs[90,143]. In adult muscle tissue, satellite cells are mononucleated muscle stem cells that can be stimulated to proliferate and differentiate to drive muscle repair[111]. Expression of the satellite cell marker gene *Pax7* decreases as satellite cells are activated into proliferating myoblasts, while expression of myogenic regulatory factors (MRFs) such as *Myod1* and *Myog* increase and promote myogenesis[111]. Satellite cells undergo asymmetric divisions to produce future *Pax7* negative, MRF positive myoblasts and to self-renew *Pax7* positive, MRF-negative satellites[144]. In addition to major transcriptional changes during myogenesis, C2C12 differentiation exhibits substantial changes, both qualitative and quantitative, in splice isoforms[145]. For example, *Pkm* undergoes an isoform switch during C2C12 differentiation that results in two distinct isozymes of the gene, PMK2 and PKM1, which include mutually exclusive exons 9 and 10 respectively[146]. Proliferating C2C12s express both isoforms of beta-tropomyosin (*Tpm2*), including exon 6a or exon 6b, but expression of the 6b isoform increases substantially during differentiation[146].

Here, we combine combinatorial barcoding of individual C2C12 cells and nuclei using the Split-seq strategy[88] with long-read sequencing (LR-Split-seq) to investigate isoform changes during differentiation. We first examined the technical differences between LR-Split-seq random hexamer and oligo-dT priming strategies as well between single-cell and single-nucleus. We compared the performance of LR-Split-seq to bulk long-read RNA-seq, and further compared the clusters recovered from LR-Split-seq to those from short-read sequencing for the same cells, as well as a companion dataset of 37,000 cells to show that long-read single-cell transcriptomes produce similar results to short-read that can be readily integrated. We then leveraged LR-Split-seq results to identify and quantify TSSs in order to perform differential TSS testing and examine TSS usage between single-cell clusters. Finally, we integrated the resulting TSS expression from LR-Split-seq with matching single-cell ATAC-seq to quantify the extent of coordinated single-cell chromatin accessibility.

2.3 Results

Comparing oligo-dT versus random hexamer primed long-read data

Split-seq uses a combination of oligo-dT and random hexamer primers in order to decrease the 3' bias that dominates other single cell RNA-seq methods that prime only with oligo-dT[88]. These methods are designed to perform 3' end counting for sequenced genes but they give little or no information about the rest of the transcript. In contrast, when Split-seq is conventionally performed with short reads, the random priming feature should, in the ideal instance, provide comprehensive information about the entire body of the transcript. However, this benefit in the short-read format is expected to have a different and unfavorable effect in long-read data. The extent and character of effects from internal priming will depend on multiple protocol variables (e.g., relative amounts of oligo-dT versus random hexamers, substrate RNA integrity) and on filtering steps in the subsequent informatic

pipeline. We therefore began by testing the impact of priming strategy on the LR-Split-seq data. We collected proliferating C2C12 myoblasts (0 h) as both whole cells and nuclei, then differentiated the remainder into myotubes over 3 days to recover 72-h differentiated nuclei (Methods). We labeled a total of approximately 37,000 cells/nuclei from the three samples using the Split-seq combinatorial barcoding strategy. We then built a sublibrary of 1000 cells for sequencing by PacBio as well as Illumina (Fig. 2.1a). The LR-Split-seq data was first debarcoded and demultiplexed using our LR-splitpipe pipeline (Fig. 2.5a) (Methods).

We then analyzed the reads with TALON[147], which is designed to assign long reads to their transcripts of origin and to identify new transcripts (Fig. 2.5b-c, Table S1) (Methods). TALON's long-read RNA-seq annotation then assigns each read to a category that specifies whether the read matches a known transcript in the reference transcriptome GTF file, or if it represents a novel transcript[147,148]. Novel reads and transcripts are further broken down by how they are novel compared to the reference annotation. Incomplete splice match (ISM) transcripts contain a subsection of an annotated transcript but do not extend all the way to the annotated 3' or 5' end. Novel in catalog transcripts (NIC) contain a new combination of exons that are all present in the reference annotation. Novel not in catalog transcripts (NNC) contain at least one splice site that is not present in the reference annotation. Antisense transcripts come from the opposite strand of a gene, and intergenic transcripts are from regions of the genome with no annotated genes. Finally, genomic transcripts overlap genes but do not share any known splice junctions with those in the annotation. Genomic transcripts are often monoexonic, short, or contain intronic regions.

Random hexamer priming is expected to start within the body of a transcript rather than the 3' polyA tail where oligo-dT primers hybridize, though intronic A-rich runs are known to serve as additional start points for oligo-dT priming[149]. This mixed priming strategy, as it is currently implemented in the Split-seq commercial platform, produced remarkably little difference in the final LR-Split-seq read length distribution from the two primer types

(Fig. 2.1b, 2.1c). The distribution of reads per TALON category showed a slightly higher proportion of incomplete splice match (ISM) reads per cell from the random hexamer priming strategy versus the oligo-dT priming strategy (Fig. 2.1c). We speculate that the high fraction of oligo-dT primed reads per cell that begin at internal sites (∼60%) accounts for the overall similarity of random hexamer primed reads in length profiles and genes detected.

Single nuclei compared with single cells for LR-split-seq

We compared single-cell versus single-nucleus LR-split-seq. Overall, more reads and genes were recovered from whole cells versus nuclei for both long- and short-read data, which is expected because cytoplasmic transcripts are left behind during nuclear extraction, making the nuclei less sensitive than whole cells (Fig. 2.1d, 2.1e). When comparing only 0 h cells with companion nuclei, we observe shorter read lengths in the nuclei (Fig. 2.1f). And as expected, we also see a larger proportion of genomic reads per cell/nucleus in nuclei compared to cells (Fig. 2.1g). These nuclear genomic reads could result from the enrichment of intronic RNA in the nucleus which would explain the lack of splice junctions.

Comparing LR-Split-seq of whole cells with bulk long-read RNA-seq for myoblasts, we found that the LR-Split-Seq is modestly shorter than bulk long-read data (Fig. 2.1h, Table S2). Bulk reads have an average mean length of 2274 bp and a peak from the kernel density estimate (KDE) distribution of 1875 bp, versus an average mean length of 1735 bp and a KDE peak of 1791 bp for LR-Split-seq non-genomic reads from whole cells (Fig. 2.1h, Table S2). The LR-Split-seq reads also had more genomic and incomplete splice match (ISM) reads than the bulk data (Fig. 2.1i, 2.1j). These differences are in line with expectations, given other differences in details of the bulk protocol (Methods). Nevertheless, after strictly filtering our novel transcripts with TALON, we retain 40,982 of the original 466,078 originally identified isoforms which represent 34.8% of reads and 34.5% of UMIs. The majority of transcript models are known transcripts annotated in GENCODE (Methods, Fig. 2.5d, Fig. 2.1k). The observed read length differences between LR-Split-seq and bulk is reflected in the

genes and transcripts that are uniquely detected in the bulk or LR-Split-seq. Transcripts detected only in bulk transcriptomes were likely to be longer, whereas transcripts detected only in LR-Split-seq data were enriched for shorter length (Fig. 2.5e, 2.5f). Due to overall longer read length in bulk long reads, these data were more likely to have multiple exons than LR-Split-seq (Fig. 2.5g). We conclude that the read length profile of known reads in single-cell LR-Split-seq is quite similar to bulk long reads, given protocol differences. This suggests to us that the overall shorter lengths in single-nucleus versus whole cell LR-split-seq are of biological origin, likely driven by underlying differences between cytosolic RNA, which is rich in mature mRNA versus nuclear RNA, which contains mature mRNA but in lower proportions.

LR-Split-seq and bulk long-read RNA-seq detect similar gene sets

Despite differences in transcript length and novelty classification between bulk long-read RNA-seq and LR-Split-seq, we detected 9584 known genes in both bulk and single-cell LR-Split-seq, with 5195 of these shared across all assays and sample combinations (Fig. 2.2a). These results demonstrate the gene detection sensitivity of LR-Split-seq. The next largest intersections contain >1500 genes recovered in all but the single-nucleus data which is likely due to the relative loss of cytoplasmic transcripts from the nuclear preparation. Genes detected in LR-Split-seq but not in the companion bulk RNA-seq tend to be short and are enriched for short RNA biotypes such as snoRNAs and miRNAs, while genes detected solely in bulk data are enriched for protein coding genes (Table S2). A plausible explanation is that Split-seq's random hexamer priming captured these transcript types whereas the bulk method, which uses oligo-dT priming exclusively, preferentially captured polyadenylated transcripts. We also examined the overlap between filtered novel transcript models from the known, NIC and NNC novelty categories in bulk and LR-Split-seq (Fig. 2.2b). While the vast majority of novel transcript models were only reproducible between bulk replicates, 251 NIC transcripts and 61 NNC transcripts were reproducible in at least one bulk and

one LR-Split-seq sample (Fig. 2.5h, 2.5I). These represent isoforms that are most likely to be real, though not previously annotated. Assuming that only the novel NIC and NNC transcripts found in bulk are real gives us a true positive rate (TPR) of 0.79 for NIC and 0.59 for NNC. We note that the calculated TPR for NIC based on the bulk is higher than for known transcripts detected by LR-Split-seq (0.71) and that it is certainly possible that an additional subset of the NIC/NNC isoforms discovered in LR-split-seq were missed in the bulk because they are lowly expressed.

LR-Split-seq recapitulates cell classifications recovered from short-read Split-seq

Overall, we recovered 110 0 h myoblast cells, 145 0 h myoblast nuclei, and 209 72 h differentiating nuclei (464 cells total) that passed short-read QC thresholds as well as an additional requirement of $\geq$ 500 long reads per cell in the 1000-cell library (Methods) (Fig. 2.6a-e). Leiden clustering based on short-read sequencing of the 464 cells/nuclei yielded 7 clusters (SR1-SR7). We observed mixed populations of 0 h myoblast cells and nuclei in clusters SR1-SR3, while the 72 h differentiating nuclei clustered in SR4-SR7. This overall structure is consistent with differentiation playing a dominant role in the UMAP structure, while differences between nuclei versus whole cells from the 0 h samples were minor by comparison (Fig. 2.2c). Additional patterns in the dataset that agree with known biology in the system include expression of the satellite cell marker gene *Pax7*, which is expressed mainly in 72 h clusters SR4 and SR5, while the key myogenic transcription factor Myog (myogenin) is expressed mainly in 72 h clusters SR6 and SR7 (Fig. 2.7a). An independent Leiden clustering performed using the LR-Split-seq data for the same 464 cells proved very similar to the companion short-read clustering with 7 clusters (LR1-LR7) in which the myoblast progenitor cells/nuclei are in clusters LR1-LR3 while the differentiating sample gives rise to clusters LR4-LR7 (Fig. 2.2d). This UMAP again separates the latter group into $Pax7^{hi}$ (LR4, LR5) and contrasting $Myog^{hi}$ sets (LR6, LR7), with the latter expressing additional downstream markers of myocyte differentiation. Color-coding cells in the long-read UMAP according to

the cluster identity from the companion short-read data showed high concordance of clusters LR4-LR7 with SR4-SR7 (Fig. 2.2e). The myoblast progenitor clusters SR1-SR3 and LR1-LR3 also agree, although the short-read clusters were more mixed between 0 h cells and nuclei.

We furthermore compared our short- and long-read cells using independent RNA velocity analyses using Velocyto[149]. We found when comparing the ratio of spliced to unspliced reads in both read formats that there was typically a higher proportion of spliced reads detected in the short reads per cell versus the long reads, which may be due to the overall higher probability of sequencing an intronic region per read in long reads (Fig. 2.7b). However, the difference is minor and the resulting independent trajectories are very similar between the short and long read experiments (Fig. 2.7c). We investigated gene expression patterns for additional known marker genes across the cells and nuclei between the short-read and long-read clusters (Fig. 2.2f). Most notably, *Mybph*, *Myh3*, and *Mef2c* are highly expressed in a subset of 72 h nuclei that make up cluster LR7, whereas *Myog* is expressed in both clusters LR6 and LR7 of 72 h nuclei (Fig. 2.2f, Fig. 2.7a). Similar to the short-read data, *Pax7* is present in both 0 h and 72 h clusters, but it is most highly expressed in clusters LR4 and LR5 (Fig. 2.2f). We also capture similar expression patterns in short-read and long-read *Pax7*^{hi} 72 h subclusters as indicated by *Igfbp5*, *Col3a1*, and *Col1a1* (Fig. 2.2f, Fig. 2.7a). Due to the consistent expression patterns of known marker genes across both technologies, we postulate that *Myog*^{hi} clusters SR6, SR7, LR6, and LR7 are mainly nuclei originating from fused, multinucleated myotubes or mononucleated myocytes on their way toward fusion, while the *Pax7*^{hi} clusters SR4, SR5, LR4, and LR5 are nuclei distinct from both myoblasts and the 72 h *Myog*^{hi} nuclei.

We examined the isoform complexity of each cell by counting the number of genes that express multiple isoforms from the same single cell or nucleus. Only one isoform was typically detected per gene in each cell. The number of genes expressing more than one isoform is

a linear function of read depth per cell, suggesting that deeper sequencing will increase isoform complexity per gene (Fig. 2.7d). Furthermore, we noticed a clear difference in the relationship between isoform complexity and read depth when comparing the single cells with single nuclei where the nuclei of increasing read depth do not display a similar large increase in isoform complexity as do the cells (Fig. 2.7d). One explanation for this is that LR-Split-seq of a nucleus captures a snapshot of its immediate state of splicing, whereas LR-Split-seq in cells captures the sum of different isoforms produced and exported to the cytoplasm over a longer period of time. If correct, the implication is that splicing within the nucleus is transiently biased for one pattern, and conceivably for one allele, the identity of which changes dynamically.

We additionally performed isoform switching tests across three identified groups of clusters: 0 h myoblast (MB) cells (LR1-LR3), 72 h *Pax7*^{hi} nuclei (LR4-LR5), and 72 h *Myog*^{hi} nuclei (LR6-LR7), with a corrected p-value cutoff from a chi-squared test of 0.05 and a change in percent isoform usage cutoff of $\geq 10\%$[133] (Methods). We recovered statistically significant isoform switching genes that have been previously observed in differentiating C2C12s, such as *Tpm2* (Adj. P = 1.06 x10^{-5} MB vs. 72 h *Myog*^{hi}) and Pkm (Adj. P = 2.57 x10^{-11} MB vs. 72 h *Pax7*^{hi}; Adj. P = 2.98 x10^{-7} MB vs. 72 h *Myog*^{hi}). The *Tpm2* locus specifically shows an increase in expression of and preference for isoforms containing exon 6b in the differentiated nuclei as previously characterized in C2C12s as visualized with Swan (Fig. 2.8a)[146,150]. We detect distinct *Pkm* isoforms with mutually exclusive exons 9 and 10 that correspond to the isozymes PKM1 and PKM2. The myoblasts tend to produce the exon 10-containing isoform (Pkm-201) over the major exon 9-containing isoform (Pkm-202), whereas the differentiated nuclei seem to equally produce Pkm-201 and Pkm-205, which has an alternative TES (Fig. 2.8b). We found 21 significant isoform-switching genes between MB nuclei and 72 h *Pax7*^{hi} nuclei as well as 14 significant isoform-switching genes between MB nuclei and 72 h *Myog*^{hi} nuclei (Table S3, Table S4).

C2C12s have distinct *Pax7^{hi}* subpopulations following differentiation

We confirmed the presence of distinct *Pax7^{hi}* and *Myog^{hi}* clusters by short-read sequencing of an extended set of cells and nuclei from the same labeled pool, comprised of six additional 9000-cell sublibraries on top of the 1000-cell sublibrary with matching long reads (Methods). After filtering, we recovered 36,869 total cells/nuclei from all seven sublibraries, including the 464 cells/nuclei with both short and long reads (Fig. 2.6a-e). The 7797 myoblast cells, 10,194 myoblast nuclei, and the 18,878 differentiating condition nuclei clustered primarily by differentiation state (Fig. 2.3a). The progenitor states form one main group in UMAP space that slightly separates cells and nuclei, while the differentiating nuclei extend outward in a spectrum with several smaller groups (Fig. 2.3a). Of the 20 clusters identified by Leiden clustering, 7 consist mostly of myoblast cells/nuclei while 13 are mainly differentiating nuclei (Fig. 2.3a) (Methods). Out of the 13 72 h clusters, 8 are *Pax7^{hi}* and the other 5 are *Myog^{hi}*, which is consistent with results from the 464 cells alone (Fig. 2.3a). Accordingly, cells from each of the 20 clusters are represented by both short and long reads in the 464-cell subset (Fig. 2.3b, 2.3c). We assign these clusters to the cells we recovered with long reads to better inform the cellular identities with high resolution (Fig. 2.9a). For example, a small subset of 12 cells out of 105 total cells in cluster SR5 belong to cluster R12, which is distinguished by high expression of *Col1a1* (Fig. 2.3d). Genes critical for cell cycle phases G1 and S such as *Cdk2* and *Pcna* are highly expressed in MB cluster R1, while G2 and M phase marker gene *Top2a* is highly expressed in MB clusters R2 (made up of mostly cells) and R3 (made up of mostly nuclei) as well as *Pax7^{hi}* cluster R9 (Fig. 2.9b)[151–153]. *Myog* is expressed throughout multinucleated myotubes as well as in some mononucleated cells that are likely to be pre-fusion myocytes (Fig. 2.9c). *Myog* and myogenic marker gene *Mybph* are highly expressed in clusters R16, R17, R18, and R20, indicating that these nuclei most likely belong to committed myocytes and myotubes (Fig. 2.3d). RNA velocity analysis, which uses the ratio of intronic (unspliced) and exonic (spliced) reads to predict the transcriptional trajectory of cells, reveals a lineage from clusters R17 and R18 toward clusters R19 and R20.

R19 and R20 express terminal myogenic marker genes such as *Myh3*, *Mef2c*, *Tnnt2*, and *Neb* (Fig. 2.3d, Fig. 2.9d) (Methods). Of the 8 co-adjacent *Pax7*^{hi} clusters (R8, R9, R10, R11, R12, R13, R14, and R15), some also express cluster-discriminating genes such as *Igfbp5* (cluster R11), *Col1a1* (cluster R12), and *Itm2a* (cluster R14) (Fig. 2.3d, Fig. 2.9d). We validated differential cluster-specificity of marker genes using spatial transcriptomic profiling of *Col1a1* (cluster R12), *Itm2a* (cluster R14) and *Myh3* (cluster R20), which showed patterns fully consistent with the Split-seq data (Fig. 2.3e). Imaging also confirmed that *Pax7*^{hi} subcluster marker genes are expressed in MNCs rather than in the multinucleated myotubes that they surround (Fig. 2.3e). *Myh3* is expressed throughout multinucleated myotubes but less so in mononucleated cells. *Pax7*^{hi} MNCs appear to either express *Col1a1* or *Itm2a*, consistent with their mutually exclusive marking of clusters R12 and R14 (Fig. 2.3d, 2.3e).

We observed heterogeneous populations of differentiating cells representing cell populations and states that are involved in adult muscle tissue repair. Clusters R10 and R11 express *Igfbp5*, which promotes muscle differentiation, and *Nfix*, which controls timing of regeneration by repressing myostatin (Fig. 2.3d, Fig. 2.9d, Table S5)[154,155]. Cluster R12, marked by *Col1a1*, *Fn1* (fibronectin), and a number of other collagen genes, may represent a population of previously defined MNCs that can transiently remodel their ECM, which is a process shown to regulate satellite cell numbers in vivo (Fig. 2.3d, Fig. 2.9d, Table S5)[156,157]. Cluster R13 expresses *Lix1*, a *Pax7* target gene needed for activated satellite cell proliferation (Fig. 2.9d, Table S5)[158]. Cluster R14, which expresses *Itm2a* and *Pax7*, may be analogous to activated satellite cells (Fig. 2.3d, Fig. 2.9d, Table S5)[159]. Appropriately, the cluster R14 RNA trajectory tends toward cluster R15 which expresses *Tead1* (*Tef-1*) and *Myog*, which are known to promote muscle differentiation (Fig. 2.9e, Table S5)[160].

Chromatin accessibility of myogenic marker genes distinguishes *Myog*^{hi} and *Pax7*^{hi} 72 h nuclei

To assess chromatin accessibility in the groups of nuclei we identified with LR-Split-seq, we

performed snATAC-seq on matching timepoints. We recovered 23,525 single nuclei from our snATAC-seq experiments following filtering and QC (Fig. 2.10a-b), resulting in 18 clusters from Leiden clustering: seven 0 h myoblast clusters and eleven 72 h differentiating clusters (Fig. 2.4a) (Methods). "Gene activity" in this context refers to a measure of chromatin accessibility of the gene body and 2kb upstream as a rough estimate of transcriptional activity[161]. We saw that chromatin gene activity patterns in our snATAC-seq UMAP for *Myog* is somewhat similar to scRNA-seq expression patterns, where the *Myog* locus was highly accessible in a subset of differentiated clusters (A16, A17, and A18) (Fig. 2.10c). To investigate the agreement between expression and chromatin accessibility for the same time points, 0 h and 72 h, we integrated our short-read Split-seq and snATAC single-cell measurements using Signac (Methods)[104]. This integration mapped Split-seq cells on snATAC-seq nuclei, resulting in predicted snATAC-seq cell types. The predicted Split-seq time point (0 h or 72 h) was mostly accurate, with 96% (10,136 out of 10,508) of true snATAC 0 h nuclei predicted to be 0 h from the expression data and 79% (10,381 out of 13,017) of true snATAC 72 h nuclei predicted as 72 h (Fig. 2.10d). When we mapped Split-seq cells grouped by MB (R1-R7), *Pax7*^{hi} (R8-R15), and *Myog*^{hi} (R16-R20) onto snATAC nuclei, we found that 48% (1502 out of 3148) of nuclei with a Myog activity score >0 were predicted to be *Myog*^{hi} and that 27% (5135 out of 18,542) of nuclei with a *Pax7* activity score >0 were predicted to be *Pax7*^{hi} (Fig. 2.10d). Unlike our Split-seq RNA data, where we detected high expression of *Pax7* in specific clusters, ATAC-based gene activity scores predicted that *Pax7* would be equally active across all clusters (Fig. 2.10e). Taken at face value, this suggests that some differentially expressed genes do not exhibit corresponding changes in promoter chromatin state, as reflected by these activity scores. However, there are several distal peaks ATAC peaks located downstream of *Pax7* whose dynamics are coordinated with the RNA. This suggests, as a working model, that they are regulatory elements governing *Pax7* expression. In contrast, *Myog* and *Mybph* illustrate expected coordinated changes in chromatin accessibility and RNA isoform expression during differentiation (clusters A16-A18) at the TSSs of

these genes (Fig. 2.10d). For uniform terminology between RNA and DNA data, we label 72 h *Myog*^{low} snATAC clusters A8-A15 as *Pax7*^{hi}. While snATAC can clearly capture changes in chromatin remodeling, the ATAC-only gene activity scores (at least as computed by Signac) do not reflect the *Pax7* expression level changes that we measure in this system.

As expected, investigation of marker peaks for *Myog*^{hi} clusters A16-A18, using a gene annotation method with gene ontology analysis, revealed significant terms such as muscle system process ($P = 1.55 \times 10^{-115}$), muscle structure development ($P = 5.77 \times 10^{-118}$), and striated muscle contraction ($P = 3.87 \times 10^{-96}$) (Methods, Fig. 2.10g, Table S6, Table S7). In comparison, MB clusters A1-A7 had broad significant terms such as regulation of anatomical structure morphogenesis ($P = 1.69 \times 10^{-19}$), cell-cell adhesion ($P = 3.45 \times 10^{-13}$), and cell motility ($P = 3.89 \times 10^{-14}$) (Table S6, Table S7). The significant terms for *Pax7*^{hi} clusters A8-A15, in contrast to *Myog*^{hi} clusters, were extracellular matrix organization ($P = 1.23 \times 10^{-9}$), extracellular structure organization ($P = 1.35 \times 10^{-9}$), and blood vessel morphogenesis ($P = 3.92 \times 10^{-9}$) (Table S6, Table S7). Most marker peaks defining the *Myog*^{hi} clusters are specific to skeletal muscle myogenesis in myotubes while marker peaks for *Pax7*^{hi} clusters indicate that they have a supportive role during development, such as by providing structural integrity to myotubes through ECM remodeling. Interestingly, cluster A9 by itself displays terms that are related to neuromuscular junction formation such as axonogenesis ($P = 0.0003$) and generation of neurons ($P = 0.002$), which indicates that a subset of differentiated nuclei might play a specialized role compared to the rest of the differentiated population (Fig. 2.10h).

LR-Split-seq identifies differential TSS choice

We developed a peak calling script to identify TSSs and TESs from long-read data (Methods). For both bulk and single-cell data, reads filtered by known, NIC, NNC, and prefix ISMs for TSSs or suffix ISMs for TESs were scanned with a window of 50bp to call TSS and TES peaks. Each end was required to be supported by at least 2 long reads (Fig. 2.11a). We

further filtered the ends at the level of each gene to achieve a refined set of TSSs and TESs for the bulk and LR-Split-seq data separately: 22,938 TSSs in bulk (Fig. 2.11b, 2.11c, Table S10), 23,996 TSSs in LR-Split-seq (Fig. 2.4b, Table S8), 14,120 TESs in bulk (Fig. 2.11d, 2.11e, Table S11), and 12,521 TESs in LR-Split-seq (Fig. 2.11f, 2.11g, Methods, Table S9). We performed the same complexity analysis on the identified TSSs and TESs per gene per cell that we did on the isoform level. We found nearly identical results where the cells and nuclei with more reads have a higher number of genes that express more than one TSS or TES and that the cells exhibit more TSS or TES complexity overall (Fig. 2.11h). Comparing the number of distinct ends to the number of distinct splice isoforms revealed that multiple TSSs are expressed per single splice isoform in both bulk and single cells (Fig. 2.11b, Fig. 2.4b). *Tnnt2* (troponin T2) has multiple known isoforms[162] and is differentially expressed between *Myog*^{hi} and *Pax7*^{hi} nuclei in the short-read data, so we decided to investigate chromatin accessibility and TSS usage at the *Tnnt2* locus (Fig. 2.9d, Table S5, Fig. 2.4c). We recovered four distinct TSSs for *Tnnt2*, three of which (Tnnt2_2, Tnnt2_3, and Tnnt2_4) overlap snATAC pseudobulk peaks, and all four of which overlap prior CAGE peaks found in C2C12[163]. Tnnt2_4 overlaps a known promoter cCRE and GENCODE vM21 transcript start site, while Tnnt2_2 overlaps a distal enhancer cCRE (Fig. 2.4c)[164]. Tnnt2_4 has both higher expression in the LR-Split-seq data and increased accessibility in snATAC *Myog*^{hi} and *Pax7*^{hi} clusters, while Tnnt2_2 and Tnnt2_3 are more highly expressed and accessible in MB clusters (Fig. 2.11i-j, Fig. 2.12a). Therefore, an isoform switch occurs in Tnnt2 where *Myog*^{hi} and *Pax7*^{hi} nuclei mainly use the known TSS belonging to the longer isoform, while the MB nuclei mainly use TSSs belonging to shorter isoforms.

Genome-wide, we validated our TSS calls using an extended set of data: our snATAC pseudobulk peaks, GENCODE vM21 TSSs, ENCODE cCREs (promoter and proximal enhancer) from mm10, and C2C12 CAGE peaks, and found that the majority of the TSSs identified from LR-Split-seq are validated by at least one of these five other datasets (Methods, Fig. 2.4d).

Using the same strategy we implemented to detect isoform switching genes, we performed differential TSS usage tests on our LR-Split-seq data (Methods). We again subset our LR-Split-seq data into 0 h MB nuclei, 72 h *Pax7*^{hi} nuclei, and 72 h *Myog*^{hi} nuclei, and performed pairwise tests. In the MB vs. *Pax7*^{hi} comparison, we found 42 genes with differential TSS usage (Table S12). In the MB vs. *Myog*^{hi} comparison, we found 40 genes with differential TSS usage. Consistent with our previous findings, this list includes *Tnnt2* (Adj. P = 6.16×10^{-14} MB vs. 72 h *Myog*^{hi}), where the MB nuclei only express isoforms consistent with downstream TSSs (Tnnt2_1, Tnnt2_2, Tnnt2_3) (Table S13). Conversely, the *Myog*^{hi} subset predominantly expresses isoforms using the upstream TSS (Tnnt2_4) (Fig. 2.4c, 2.4e).

Similarly, we found multiple distinct TESs per splice isoform in bulk and LR-Split-seq data (Fig. 2.11d, 2.11f). We validated genome-wide TESs using GENCODE vM21 TESs and polyA-seq peaks from C2C12 cells at days 0 and 4 of differentiation, which overlapped the majority of TESs found in bulk data but not in the LR-Split-seq data (Fig. 2.11e, 2.11g) (Methods). We believe that the apparent lack of external validation for the LR-Split-seq TESs is largely driven by the random priming method. When we call TESs instead using the same set of reads but split by priming strategy, 50.3% of the oligo-dT TESs validate by at least one form of external support compared to 6% of the random hexamer TESs (Fig. 2.12b). When we use the same strategy to compare TSSs from oligo-dT-primed reads to those from random hexamer-primed reads, 83.5% of the oligo-dT TSSs validate by at least one form of external support compared to 84.8% of the random hexamer TSSs (Fig. 2.12c, 2.12d).

We furthermore demonstrated the utility of LR-Split-seq for identifying TSSs and TESs by comparing how well our long-read ends are supported by external validation in comparison to those we called with from our companion short-read Split-seq data for the same cells. Only 50.2% of TSSs called using the short reads had external validation compared to 81.5% of LR-Split-seq TSSs (Fig. 2.4d, Fig. 2.12e) (Methods). Similarly, the short-read Split-seq

TESs validate externally at a much lower rate (12.3%) than the LR-Split-seq TESs (44.2%) (Fig. 2.11g, Fig. 2.12f) (Methods).

Coordination of chromatin accessibility with transcriptional output

We calculated snATAC TSS chromatin accessibility across our refined set of TSSs to determine how TSS accessibility relates to TSS expression. We found that binary chromatin accessibility at a TSS was a particularly strong indicator of whether or not a TSS was expressed in MBs (65.7% of TSSs) (Fig. 2.12g) and that the level of accessibility at each TSS correlated well with the expression level of each TSS (Pearson r = 0.44, Spearman rho = 0.58) (Methods). In the *Pax7*^{hi} and *Myog*^{hi} populations, accessibility at each TSS did not correlate as strongly (Pearson r = 0.58, Spearman rho = 0.20; Pearson r = 0.53, Spearman rho = 0.16 respectively). We then determined how many genes with more than one TSS displayed the highest accessibility level and expression level at the same TSS. In the myoblasts, the most highly-accessible TSS for a gene was most often also the most highly-expressed TSS for the gene (77.8% of genes with >1 TSS). This concordance was less strong in the *Pax7*^{hi} and *Myog*^{hi} groups (54.6% and 52.7% respectively).

We then investigated which TSSs are supported by both differential accessibility and expression (Methods). We compared the average log2 fold change (LFC) in both accessibility and expression between *Myog*^{hi} and MB (Fig. 2.4f), *Pax7*^{hi} and MB (Fig. 2.4g), and *Myog*^{hi} and *Pax7*^{hi} (Fig. 2.4h). Between MB and *Myog*^{hi}, 19 TSSs are specific to *Myog*^{hi} with an average LFC greater than two standard deviations (indicated by dashed lines) in both datasets, and 70 TSSs are specific to MB with average LFC less than two standard deviations in both datasets (Fig. 2.4f, Table S14). Several of the genes with such TSSs are differentially expressed (Table S5, Table S14). Only 6 TSSs were *Pax7*^{hi}-specific relative to MB, but one of these is *Igfbp5*, which is a gene that was highly differentially expressed in the *Pax7*^{hi} subset (Fig. 2.4g, Fig. 2.9d, Table S3, Table S14). Comparing MB and *Pax7*^{hi}, 77 TSSs are MB-specific, 36 of which are also MB-specific when comparing *Myog*^{hi} with MB. Of the 19

Myog^{hi}-specific TSSs between *Myog*^{hi} and MB, 15 were also *Myog*^{hi}-specific when compared to *Pax7*^{hi} (out of 53 total) (Fig. 2.4h, Table S14). Several of the 17 *Pax7*^{hi}-specific TSSs (Fig. 2.4h) belong to differentially expressed genes, such as *Pax7*, *Col4a1*, *Fn1*, and *Igfbp5* (Fig. 2.9d, Table S5, Table S14). From a biological perspective, *Prox1* and *Vgll4* are potentially interesting; although they were not differentially expressed in the short-read data, they are known to be involved in skeletal muscle regeneration (Fig. 2.4h, Table S14)[165,166].

2.4 Discussion

The first goal of this work was to advance our capacity to directly map and quantify RNA isoforms in single cells. Using the C2C12 myogenic differentiation as a test system, we introduce long read-Split-seq (LR-Split-seq) and show that it can be as effective as standard short-read Split-seq for detecting cell clusters, based on data from the same number of cells or nuclei. This conclusion applied to nuclei as well as whole cells, although whole-cell data detected more genes per cell than companion LR-Split-seq data from nuclei. For biological systems that do not permit uniform whole-cell disaggregation such as our multinucleated myotubes or brain tissue, the success shown here for nuclei is encouraging. We speculate that the remaining sensitivity differential between nuclei and whole cells is a consequence of the smaller starting number of transcripts in nuclei, and some of that could be further compensated by increasing the nuclear number sequenced and their depth of sequencing. We also suggest that combining random hexamer primed long-reads with the oligo-dT primed long-read data helped to capture 5' ends that are critical for inferring TSS use, although this adds incomplete PacBio reads to the overall dataset. We also illustrate that LR-Split-seq affords users the choice of analyzing the oligo-dT primed and hexamer primed read populations separately. A second motive for developing LR-Split-seq is that it will allow flexible study designs that can efficiently and more economically refine cell type identities

by integrating additional standard short-read Split-seq data on the same samples. Results presented here showed that this strategy was effective in refining stem cell identities and states in the C2C12 system. Finally, we integrate results from LR-Split-seq with snATAC to gain insights into the dynamics of chromatin accessibility at the corresponding promoters with a longer term goal of building a fully integrated model of physically or genetically affiliated distal regulatory elements.

We were able to detect 79% of the genes and 53% of transcript isoforms detected in bulk myoblast long-read RNA-seq using LR-Split-seq in single cells. We expect these differences relative to bulk samples to be a function of the individual study design, including number and diversity of cells sequenced, depth of sequencing, fixation protocol and, for isoform detection, the contribution from internal hexamer priming. The largest sets of genes detected across the entire analysis included the LR-Split-seq assays, supporting the conclusion that it detects expressed genes reliably and reproducibly. The differences between known gene and transcript detection rates, relative to bulk data, were largely attributable to internally-primed Split-seq reads and their management in our computation pipeline. Specifically, we used TALON, which leverages non-full-length reads for quantification and detection on the gene-level but not on the transcript-level. Consequently, we achieved high gene detection concordance but lower transcript detection concordance between long-read bulk and LR-Split-seq data.

Gene-level clusters in LR-Split-seq are remarkably similar to the results in the equivalent standard short-read Split-seq. In both assays, clusters of differentiating cells were most homologous to each other and were distinct from the myoblast clusters. However, in LR-Split-seq, there was a greater tendency for the clusters to separate by assay format, as shown in the 0 h myoblast cells and nuclei. We captured expression dynamics of well-known myogenic marker genes in the differentiating clusters such as *Pax7*, *Myog*, *Mybph*, and *Myh3* that are reproducible in the short-read data we sequenced from the same cells[111].

The additional context from ∼37,000 short-read single cells allowed us to investigate the *Myog*^{hi} clusters in greater detail. We found that *Myog*^{hi} clusters were very distinct from MB clusters, while *Pax7*^{hi} clusters were in a spectrum of differentiation stages between MB and *Myog*^{hi} clusters. Expression of additional marker genes in *Pax7*^{hi} subclusters, RNA velocity trajectories, and validation with spatial transcriptomic profiling confirmed that these nuclei are from mononucleated cells in varying stages of differentiation.

LR-Split-seq enabled us to investigate transcript-level differences between the various stages of differentiation in myogenesis. We found novel insights into the biology of the system by studying differential TSS usage and integrating our TSSs identified from long reads and our snATAC-seq peaks. Our analysis revealed over 50 significant switches in TSS usage across clusters of undifferentiated versus differentiated stages, including a pronounced switch in *Tnnt2*, where the myoblasts primarily use TSSs that are novel to more recent GENCODE transcript annotations, while differentiated cells mainly express the known TSS that results in a longer isoform. This TSS switch was complemented by a corresponding increase in chromatin accessibility at the newly-expressed TSS in *Myog*^{hi} clusters.

Unlike previous long-read scRNA-seq methods that rely on sequencing of each cell using custom microfluidics equipment[131,133], LR-Split-seq is immediately accessible with no cell/droplet handling instrumentation and it is tunable in both cell number and sequencing depth, depending on the complexity of the underlying sample's cellular composition. Additionally, it can be scaled up for long-read sequencing with additional sublibraries and higher read depth. We believe that this will allow one to optimize the amount and character of information from short and long-read single-cell technologies when the costs of input cells, overall platform, and sequencing are all considered. While short-read Split-seq provides a broad survey of the transcriptional complexity of a biological system by sequencing up to 100,000 cells, corresponding LR-Split-seq can be applied to a targeted number of cells to provide higher-resolution isoform-level insights using a few million long reads from a few PacBio

runs. In this way, LR-Split-seq promises relatively affordable, simultaneous transcriptional profiling of a wide variety of tissues using short and long-read sequencing.

Acknowledgements

We would like to thank Melanie Oakes at UC Irvine Genomics High-Throughput Facility (GHTF) for her help with PacBio sequencing.

Contributions

Gabriela Balderrama-Gutierrez, Isaryhia Rodriguez, Cassandra McGill, and Heidi Liang performed the experiments. Fairlie Reese (co-author) analyzed the data and wrote the manuscript with significant input from Katherine Williams, Barbara J. Wold, and Ali Mortazavi. Fairlie Reese supplied custom LR-Split-seq demultiplexing code and Diane Trout provided TSS-calling code. All authors read and approved the final manuscript.

Data availability

- Long read RNA-seq experiment for bulk 0hr C2C12
- Long read RNA-seq experiment for bulk 72hr C2C12
- Paired long read and short read Split-seq experiments for single-cell 0hr C2C12
- Paired long read and short read Split-seq experiments for single-nucleus 0hr C2C12
- Paired long read and short read Split-seq experiments for single-nucleus 72hr C2C12
- Short read Split-seq experiment for single-cell 0hr C2C12 (9000-cell sublibraries)
- Short read Split-seq experiment for single-nucleus 0hr C2C12 (9000-cell sublibraries)
- Short read Split-seq experiment for single-nucleus 72hr C2C12 (9000-cell sublibraries)
- ATAC-seq experiment for single-nucleus 0hr C2C12

- ATAC-seq experiment for single-nucleus 72hr C2C12
- ATAC-seq experiment for filtered single-nucleus 72hr C2C12

Code availability

- Demultiplexing and debarcoding tool designed for LR-Split-seq data
- Data processing and figure generation code

2.5 Supplementary tables

- **Table S1: TALON read annotation file for long-read data.**
- **Table S2: Gene biotype enrichment in long-read bulk vs. single-cell data.**
- **Table S3: Isoform switching in myoblasts vs. *Pax7*^{hi}.**
- **Table S4: Isoform switching in myoblasts vs. *Myog*^{hi}.**
- **Table S5: Cluster marker genes for short-read Split-seq.**
- **Table S6: Cluster marker peaks for snATAC-seq.**
- **Table S7: GO term enrichment from snATAC-seq marker peaks.**
- **Table S8: BED file of TSSs in bulk.**
- **Table S9: BED file of TSSs in LR-Split-seq.**
- **Table S10: BED file of TESs in bulk.**
- **Table S11: BED file of TESs in LR-Split-seq.**
- **Table S12: TSS switching in myoblasts vs. *Pax7*^{hi}.**

- **Table S13: TSS switching in myoblasts vs. *Myog*^{hi}.**
- **Table S14: TSS fold changes measured by RNA and ATAC data between similar comparisons (supplement to 2.4f-h).**

2.6 Methods

C2C12 culture and differentiation

C2C12 cells were purchased from the American Type Culture Collection (ATCC, CRL-1772). All cells used in experiments were passaged less than 10 times from the original plug. C2C12 were authenticated by testing for differentiation efficiency upon receipt. They were not tested for mycoplasma throughout the course of the study. C2C12 cells were cultured on 10-cm plates (Thermo Scientific, 172931) in 10 mL myoblast growth media: high-glucose DMEM with L-glutamine and without sodium pyruvate (HyClone, SH30022.FS), supplemented with 20% fetal bovine serum (Omega Scientific, FB-11), 100 units/mL penicillin, and 100 ug/mL streptomycin (Gibco, 15140122). Cells were maintained at 20-50% confluency at 37°C with 5% CO_2 and passaged at 1:3 or 1:4 every 2 to 3 days. To detach them from plates, cells were rinsed with 1X PBS (HyClone, SH30256.02) and incubated with 2 mL TrypLE-Express (Gibco, 12605010) for 5 min at 37°C, which was then neutralized with 8 mL myoblast growth media. To differentiate, cells at 90-100% confluency were rinsed with 1X PBS and myoblast growth media was replaced with 10 mL differentiation media: high-glucose DMEM with L-glutamine and without sodium pyruvate (HyClone, SH30022.FS), supplemented with 2% donor horse serum (Gibco, 16050130), 100 units/mL penicillin, 100 ug/mL streptomycin (Gibco, 15140122), and freshly-added 1 µM insulin (Sigma-Aldrich I6634). Differentiation media was replaced every 24 hours for 3 days. Cells were monitored under a microscope (EVOS FL Auto 2) to observe changes in morphology and confirm differentiation.

Preparation of myoblast and myotube single-nucleus suspensions

We followed the Bio-Rad SureCell WTA 3' Library Prep protocol for preparation of nuclei samples[167]. Myoblasts from one 10-cm plate (∼1.5 million cells) and myotubes from one 10-cm plate (∼5 million cells) with >90% viability were lifted as described above and pelleted in 15-mL polypropylene falcon tubes (VWR, 89039-670) by centrifuging for 5 min at 1500 RPM. Cells were washed twice with cold 1X PBS + 0.1% BSA (Sigma-Aldrich A9418) and 0 h myoblasts were filtered through a 40-µm strainer; due to their size, 72 h samples containing myotubes were not filtered. After centrifuging for 3 min at 300 × *g*, cells were resuspended in 1 mL cold lysis buffer: 10 mM Tris-HCl pH 8 (Thermo Scientific, AM9855G), 10 mM NaCl (Fisher Scientific, S271), 3 mM MgCl2 (Sigma, M8266), 0.1% IGEPAL CA-630 (Thermo Scientific, 28324), 0.2 U/µL SUPERase In RNase Inhibitor (Invitrogen, AM2694) and 10 mg/mL BSA in nuclease-free water (Ambion, AM9937). Cells were incubated in lysis buffer on ice for 10 min, centrifuged at 4°C for 3 min at 300 × *g*, and washed with 1 ml of cold 1X PBS + 1% DEPC water (Invitrogen, 750023). The lysis, spin, and wash steps were repeated two more times for the 72 h samples because myotube cell membranes are more difficult to fully lyse than mononucleated myoblasts. Nuclei were stained with Trypan Blue (Bio-Rad, 1450021), and cell membrane lysis was confirmed under a microscope and by percent viability (<10%). Nuclei were stored on ice in 1 mL nuclei storage buffer (lysis buffer without the addition of IGEPAL CA-630).

Preparation of single-cell barcoded cDNA using Split-seq

Single-cell barcoded cDNA and Illumina libraries were prepared using the Fixation Kit for Cells, Fixation Kit for Nuclei, and Single Cell Whole Transcriptome Kit (Parse Biosciences, SB2001) following the manufacturer's protocols. Nuclei from the 0-h myoblast sample and 72-h sample in single-nucleus suspensions were counted on a TC20 Automated Cell Counter (Bio-Rad, 1450102), and ∼4 million were filtered through a 40-µm strainer into 15-mL polypropylene falcon tubes. Nuclei were fixed for 10 min and permeabilized for 3 min on ice,

then DMSO was added for storage overnight at -80°C in a Mr. Frosty. Myoblast cells were similarly counted and filtered through a 40-µm strainer, followed by fixation and permeabilization. DMSO was added and cells were stored overnight at -80°C in a Mr. Frosty. Before storage, single-cell and single-nucleus suspensions were confirmed under a microscope.

To prepare barcoded cDNA, fixed and frozen cells and nuclei were thawed in a 37°C water bath and counted. Cells were added to the Round 1 reverse transcription barcoding plate at around ∼15,000 cells/well, with A1-A12 containing 0 h cells, B1-B12 containing 0 h nuclei, and C1-D12 containing 72 h nuclei (Fig. 2.6a), before in situ reverse transcription and annealing of barcode 1+linker on a thermocycler (Bio-Rad T100). After RT, cells were pooled using a multichannel pipette into a 15-mL tube, spun down at 4°C for 5 min at 1000 × g, and resuspended in 1 mL of Resuspension Buffer (Parse Biosciences, SB2001). Using a basin and multichannel pipette, cells were distributed in 96 wells of the Round 2 ligation barcoding plate for the in situ barcode 2+linker ligation. Next, cells were pooled, filtered through a 40 µm strainer, and redistributed into 96 wells of the Round 3 ligation barcoding plate for the in situ barcode 3+UMI+Illumina adapter ligation. After a final pooling and filtration through a 40-µm strainer, cells were counted using a hemocytometer and distributed into 7 sublibraries: 6 sublibraries with 9000 cells each, and 1 sublibrary with 1000 cells. The cells in each sublibrary were lysed and libraries were cleaned with AMPure XP beads (Beckman Coulter, A63881), then the single-cell barcoded cDNA underwent template switching and amplification. Importantly, we increased the number of cycles for the 1000-cell library to 20 cycles rather than 18 in order to increase the yield of single-cell barcoded cDNA for use in Illumina library preparation (50 ng) while having enough leftover cDNA for PacBio library preparation (500 ng). The cDNA was cleaned using AMPure XP beads and quality checked using an Agilent Bioanalyzer before proceeding to Illumina and PacBio library preparation.

Preparation of Illumina scRNA-seq libraries using Split-seq and sequencing

All 7 sublibraries were fragmented, size-selected using AMPure XP beads, and Illumina

adapters were ligated. The cDNA fragments were cleaned again using beads and amplified, adding the fourth barcode and P5/P7 adapters, followed by a final bead-based size selection and quality check with a Bioanalyzer. Libraries with 5% PhiX spike-in were loaded at 2.1 pM and sequenced to an average depth of 51 million reads per 9000-cell library and 70 million reads for the 1000 cell library using an Illumina NextSeq 500 with paired-end run configuration 74/86/6/0. The data are hosted on GEO (GSE168776) and on the ENCODE portal (ENCBS521YWL, 0 h cells, ENCBS431NOZ, 0 h nuclei; and ENCBS978ZNQ, 72 h nuclei).

Preparation of PacBio scRNA-seq library and sequencing

The PacBio library was prepared using 500 ng of amplified, single-cell barcoded cDNA with the SMRTbell Template Prep Kit (PacBio, 100-938-900) according to the manufacturer's protocol for sequencing on a Sequel II. The 1000-cell library was sequenced using 2 SMRTcells (PacBio, 101-008-000) for a sequencing depth of 5,764,421 full-length non-chimeric reads. The data are hosted on GEO (GSE168776) and on the ENCODE portal (ENCBS521YWL, 0 h cells, ENCBS431NOZ, 0 h nuclei; and ENCBS978ZNQ, 72 h nuclei).

Preparation of bulk PacBio libraries and sequencing

We extracted RNA from two replicates of C2C12 0 h samples and 72 h samples using the RNA-easy kit (Qiagen, 74104). cDNA synthesis and library preparation using the SMRTbell Template Prep Kit (PacBio, 100-938-900) were performed as described on the ENCODE portal (`https://www.encodeproject.org/documents/77db752f-abf7-4c93-a460-510464134f52`). We sequenced one SMRT cell per replicate on the Sequel II platform. The data are hosted on the ENCODE portal (ENCBS824FPY, ENCBS649CMC for 0hr cells; and ENCBS373BHL, ENCBS606QKU for 72hr cells).

Preparation of snATAC-seq libraries using Bio-Rad technology and sequencing

The single nucleus ATAC-seq experiment was performed using the SureCell ATAC-Seq Library Prep Kit (Bio-Rad, 17004620) following the manufacturer's protocol for the OMNI-ATAC version[168]. Cells at 0 h differentiation or 72 h differentiation timepoints in one 10-cm plate per biological replicate were lifted as previously described and washed twice in cold 1X PBS + 0.1% BSA. All 0 h replicates and some 72 h replicates were filtered through a 40-µm strainer (2 technical replicates, 1 biological replicate; 2 technical replicates of 72 h samples were not filtered), then counted and assessed for viability. 300,000 cells with >90% viability per biological replicate were lysed with cold OMNI-ATAC lysis buffer on ice for 3 min and washed out with cold ATAC-Tween buffer, at which point non-filtered 72 h nuclei were filtered through a 40-µm strainer, then spun down at 500 RCF for 10 min at 4°C. Nuclei were resuspended, counted, and confirmed to be single-nucleus suspensions under a microscope, then 60,000 nuclei per biological replicate were tagmented at 37°C for 30 min in a ThermoMixer with 500 RPM mixing. The microfluidics-based ddSEQ Single-Cell Isolator was used to stream tagmented nuclei in an amplification reaction mix with barcoded beads to isolate single nuclei in nanodroplets with one or more barcodes. Tagmented cDNA was barcoded and amplified, then nanodroplets were broken and libraries cleaned with AMPure XP beads before a second amplification of barcoded fragments and final bead-based cleanup. A Bioanalyzer was used to verify library quality before loading at 1.5 pM and sequencing to an average depth of 122 million reads per library using an Illumina NextSeq 500 with paired-end run configuration 118/40/8/0 and custom sequencing primer. The data are hosted on GEO (GSE168776) and on the ENCODE portal (ENCBS081AJF, ENCBS562OEW, 0hr nuclei; ENCBS779SXF, ENCBS143VME, 72hr nuclei; and ENCBS247OBN, ENCBS090IYH, 72hr nuclei isolated from filtered cells).

Validation of transcript expression with RNAscope

Myoblasts were grown to 90-100% confluency in flasks mounted on slides (Thermo Scientific, 170920) then differentiated over 3 days as previously described. The flasks were removed

and slides were rinsed in 1X PBS, followed by fixation in 10% neutral buffered formalin (Sigma-Aldrich, HT501128) for 30 min at room temperature. Following the manufacturer's protocol for cultured adherent cells, we rinsed the slides in 1X PBS, then incubated in 50%, 70%, and 100% ethanol for 5 min each[169]. Slides were stored submerged in 100% ethanol at -20°C in 50 mL falcon tubes. To rehydrate, slides were incubated in 70% and 50% ethanol for 2 min each, then in 1X PBS for 10 min. A hydrophobic barrier was drawn around the edges of the slide (Vector Laboratories, H-4000), then the cells were permeabilized with 1:15 diluted protease III (ACDBio, 322340) for 10 min at room temperature in a humidity control tray (ACDBio, 310012). Following the manufacturer's protocol for the RNAscope HiPlex12 kit (ACDBio, 324100/324140), probes for genes of interest were mixed and hybridized for 2 hours at 40°C in a HybEZ II hybridization oven (ACDBio, 321710/321720), then the signal was amplified over 3 rounds of 30 min incubations at 40°C in the oven[170]. We then proceeded to fluorophore hybridization and imaging over four rounds of three channels per round (GFP, RFP, and Cy5) plus DAPI[171]. An EVOS FL Auto 2 with programmable stage was used to automatically image slides at $\times$ 40 magnification.

Preprocessing of LR-Split-seq data

Raw PacBio reads were processed into circular consensus reads using the ccs software from the SMRT analysis software suite (parameters: `--skip-polish --min-length=10 --min-passes=3 --min-rq=0.9 --min-snr=2.5`) (`https://github.com/PacificBiosciences/ccs`). The Split-seq adapters were identified and removed using Lima (v2.0.0) (parameters: `--ccs --min-score 0 --min-end-score 0 --min-signal-increase 0 --min-score-lead 0`) (`https://github.com/pacificbiosciences/barcoding/`). Reads were then processed with IsoSeq3's Refine (v3.4.0) to yield full-length non-chimeric reads (`https://github.com/PacificBiosciences/IsoSeq`). As around half of our reads are primed using random hexamer priming, polyA tails were not required nor removed for this step.

Reads were then demultiplexed for their Split-seq barcodes using a custom script (`https://`

github.com/fairliereese/LR-splitpipe) by first detecting the spacer sequences between barcodes and using these as start and end points for the barcodes. Barcodes were corrected to those that were within an edit distance of 3 of the predetermined list of barcodes used for each round of barcoding. The resultant reads were then filtered on which combinations of barcodes were also seen in the Illumina single cell/nucleus RNA-seq data, which yielded 567 of the 568 cells that passed QC in the Illumina data (Fig. 2.6b). The reads were then trimmed of their barcodes to facilitate mapping, and cell identity barcodes were recorded. The reads were mapped using Minimap2 (v2.17-r94) (`-ax splice:hq -uf --MD`)[172] and the mm10 reference mouse genome, corrected for long-read sequencing artifacts with TranscriptClean (`--canonOnly --primaryOnly`)[173]. We then used TALON (development branch on GitHub) (`--cb`) to annotate each read to its transcript or origin using the GENCODE vM21 reference[147]. We filtered for reproducible novel NIC and NNC transcript models for those that were seen in 4 or more sub-cells (Fig. 2.1k, Fig. 2.5d). Custom LR-Split-seq demultiplexer can be found at `https://github.com/fairliereese/LR-splitpipe`[174] or on Zenodo at `https://doi.org/10.5281/zenodo.5168057`. Figure generation code can be found at `https://github.com/fairliereese/2021_c2c12`[175] or on Zenodo at `https://doi.org/10.5281/zenodo.5168059`. All code is available under the MIT open source license.

Comparing priming strategies and sample types in LR-Split-seq data

The priming strategy of each read was determined by examining the barcode for the first round of Split-seq. Reads were separated out by priming strategy and by cell. For sample comparisons, the oligo-dT and random hexamer primed reads from each cell were merged to create the final cell, then separated out by sample.

Comparing bulk long-read to LR-Split-seq

To enable this comparison, we re-ran the bulk and single-cell data through TALON using

the same database so that novel transcripts would have the same IDs across the bulk and the single-cell. For the bulk novel transcript models, filtering was done using `talon_filter_transcripts`, requiring a novel transcript model to be reproducible in at least 2 of the bulk replicates with at least 5 copies. For the single-cell, filtering was done that required novel transcript models to be reproducible in at least 4 sub-cells.

Single-cell processing of LR-Split-seq data

Oligo-dT and random hexamer primed reads from each cell were merged to create the final cells. Gene-level cells and nuclei were further filtered for those that had $\geq$ 500 reads per cell/nucleus using Scanpy (v1.4.6)[99] and for those that, in the corresponding Illumina data, had <200,000 reads, <20% mitochondrial reads, and >500 genes (done in Seurat as detailed in the Processing of short read scRNA-seq data section) (all on a per cell/nucleus basis); yielding a final total of 464 single cells and nuclei. Dimensionality reduction, construction of the UMAP, and Leiden clustering were all performed using Scanpy, yielding 7 clusters (Fig. 2.2d).

Isoform switching gene testing

Testing for isoform switching in LR-Split-seq data was performed as in Joglekar et. al., 2021[133]. For each pairwise test, an n $\times$ 2 contingency table was created with counts in each condition for each isoform in a gene, with a maximum of 11 isoforms. In cases where a gene had more than 11 isoforms, an 11th entry was constructed where counts were summed for the most lowly expressed isoforms. Each gene was required to have at least 10 supporting reads from each condition to be considered testable. For each testable gene, a chi-squared test was performed and, $\Delta\pi$, or the change in percent isoform usage for the gene, was computed as the sum of the absolute value of percent isoform usage across the conditions for the top two expressed isoforms. P values from the chi-squared test were corrected using Benjamini-Hochberg correction. Tests were performed on the LR-Split-seq Leiden clusters for MB nuclei

vs. *Myog*^{hi} nuclei (clusters LR1-LR2 vs. LR6-LR7) and for MB nuclei vs. *Pax7*^{hi} (clusters LR1-LR2 vs. LR4-LR5). Genes with significant isoform switching were required to have a corrected p value ≤ 0.05, $\Delta\pi$ of ≥ 10, and a minimum number of reads per gene per tested condition of 10.

Processing bulk long-read data

Bulk PacBio data was processed following the ENCODE Long Read RNA-Seq Analysis Protocol for Mouse Samples (v.1.0) for CCS, Lima, refine and TranscriptClean steps (`https://www.encodeproject.org/documents/a84b4146-9e2d-4121-8c0c-1b6957a13fbf`). A TALON database was initialized using mm10 GENCODE v21 GTF with SIRV set 3 and ERCCs included. Reads output from TranscriptClean were labeled with the corresponding fasta reference. TALON was run (`--cov 0.9 --identity 0.8`). Filtering novel transcript models was done using TALON's `talon_filter_transcripts` module, requiring a novel transcript model to be reproducible across biological replicates, and appear 5 times in each replicate, as well as display a lack of internal priming evidence (`--minCount 5 --minDatasets 2 --maxFracA 0.5`). Transcript abundances were determined using talon_abundance.

Processing of short read Split-seq data

After initial demultiplexing of the 7 sublibraries (6× 9000-cell sublibraries and 1 1000-cell sublibrary), Parse Bioscience's split-pipe v0.7.6 software was used to deconvolute reads into single cells, map to mm10 using STAR (v. 2.6.0c), annotate using GENCODE vM21, and filter using a UMI cutoff determined by knee plots (Fig. 2.6b, 2.6d)[176]. The remaining cells were further filtered in Seurat (v. 3.2.2) by <20% mitochondrial reads, <200,000 counts, and >500 genes per cell/nucleus (Fig. 2.6c, 2.6e)[177]. The resulting 464 cells with both short and long reads and 36,405 cells with short-read data only were analyzed using Velocyto (v.0.1.17)[149]. 55% of counts from 0 h cells, 46% of counts from 0 h nuclei, and 37% of counts from 72 h nuclei were spliced out of the total number of spliced and unspliced

counts. After loading the loom file back into Seurat with the ReadVelocity function from the SeuratWrappers package, SCTransform (v. 0.3.1) was used to regress percent mitochondrial reads, number of genes, and sublibrary, followed by UMAP dimensionality reduction[101,178]. Clustering using the Leiden algorithm (v. 0.8.0) resulted in 20 clusters[100]. Differentially expressed genes per cluster were found using Seurat's FindAllMarkers function (`only.pos = TRUE, min.pct = 0.1, logfc.threshold = 0.1`) then further filtered by FDR <0.01.

Processing of snATAC-seq data

After demultiplexing the 8 snATAC-seq libraries (3× 0 h, 5× 72 h samples), Bio-Rad's dockerized ATAC-seq analysis toolkit (v.1.0.0) was used to recover barcodes/UMIs, align reads with BWA, filter and deconvolute barcodes, perform quality control by UMI thresholding, and call peaks with MACS2 (Fig. 2.10a)[179–181]. A custom script (`https://github.com/fairliereese/lab_pipelines/tree/master/sc_atac_pipeline`) that takes in the combined peaks file, QC-passing barcode list, and mapped reads was used to generate peaks-by-cells counts matrices as csv files for each library. In addition, the annotated bam files were converted to fragment files using scATAC-pro's simply_bam2frags.pl script, which are bed-like matrices containing chromosome, start, stop, cell ID, and number of fragments contained in the region[182]. Further QC cutoffs consisted of a TSS enrichment score >6, >5,000 counts, and <20,000 counts per nucleus (Fig. 2.10b). TSS enrichment is calculated in Signac (v. 1.0.9004) following the definition by ENCODE (`https://www.encodeproject.org/data-standards/terms/`). Signac was used to normalize binarized peaks-by-cells counts matrices by term frequency inverse document frequency (TF-IDF) followed by singular value decomposition and UMAP dimensionality reduction[161]. The Leiden algorithm (v. 0.8.0) was used to resolve 18 clusters (Fig. 2.4a). UCSC Genome Browser tracks were generated by splitting the snATAC bam file by cluster using the sinto package (`https://github.com/timoast/sinto`) and creating bigwig tracks using deeptools[183]. Differentially accessible peaks per cluster were found using Seurat's FindAllMarkers func-

tion (`only.pos = TRUE, min.pct = 0.1, logfc.threshold = 0.5`) then further filtered by FDR <0.05. The marker peaks were grouped by MB (A1-A7), *Pax7*^{hi} (A8-A15), and *Myog*^{hi} (A16-A17) and processed using GREAT with mm10 whole genome background and associating peaks with the single nearest gene within 50kb[184]. *P* values for the binomial test are reported in the text. The resulting genes for cluster A9 were also input into Enrichr to determine enriched GO Biological Processes[185]. The clustergram was downloaded from the Enrichr web tool (`https://maayanlab.cloud/Enrichr/`).

Integration of short-read Split-seq and snATAC-seq data

Signac's FindTransferAnchors function was implemented with all 36,869 Split-seq cells as the reference set and all 23,525 snATAC-seq nuclei as the query set, with canonical correlation analysis (CCA) used as the dimensional reduction method[104]. The TransferData function was used to carry over Split-seq labels "0 h" or "72 h" in one analysis (Fig. 2.10d, left panel) and labels "MB," "*Pax7*^{hi}," and "*Myog*^{hi}," in another analysis (Fig. 2.10d, right panel).

Identification of TSSs from long-read data

For both LR-Split-seq and bulk separately, bam reads were filtered for those that were annotated by TALON as belonging to the known, novel in catalogue (NIC), novel not in catalogue (NNC), and prefix-ISM novelty categories as the starts of reads belonging to these novelty categories are more likely to come from a true 5' end. TSSs were called on the filtered bams using the ENCODE PacBio TSS caller (`https://github.com/ENCODE-AWG/tss-annotation/blob/master/long_read/pacbio_to_tss.py`) (`--window-size=50 --raw-counts --expression-threshold=0`), yielding a bed entry for each TSS consisting of a wide peak, narrow peak, and a summit for each TSS. Resultant Split-seq TSSs were filtered first by requiring each one to be supported by at least 2 reads, and subsequently on the gene level, where each called TSS was required to have a number of reads >10% of the number of reads that supported the most highly expressed TSS for the same gene.

Bulk TSSs were similarly filtered except using a threshold of >5%. For the oligo-dT versus random hexamer priming comparison, the aforementioned filtered reads were split based on their priming strategy before running the TSS caller. The intersection of oligo-dT TSSs and random hexamer TSSs was determined using bedtools v. 2.30.0[186].

Identification of TESs from long-read data

Similarly, for both LR-Split-seq and bulk data separately, bam reads were filtered for those annotated as belonging to the known, novel in catalogue (NIC), novel not in catalogue (NNC), and suffix-ISM novelty categories as the ends of reads belonging to these novelty categories are more likely to come from a true 3' end. TESs were called on the filtered bam using the same ENCODE PacBio TSS caller (`https://github.com/ENCODE-AWG/tss-annotation/blob/master/long_read/pacbio_to_tss.py`) (`--window-size=50 --raw-counts --expression-threshold=0 --tes`), yielding a bed file the same format as the TSS file. The TESs were then filtered by requiring each to be supported by at least 2 reads and have a number of reads >80% of the number of reads that supported the most highly-expressed TES for the same gene. For the oligo-dT versus random hexamer priming comparison, the aforementioned filtered reads were split based on their priming strategy before running the TSS caller.

Identification of TSSs from short-read data

Bam reads from the short-read Split-seq 464 cells were queried for those that contained the complete template-switching oligo sequence with no errors using Cutadapt v. 2.10 (`-g AACGCAGAGTGAATGGG -e 0 -O 17`), which represent reads that are more likely to contain a true 5' end[187]. TSSs were called on the filtered bams using the ENCODE PacBio TSS caller (which despite its name works on short reads as well) `https://github.com/ENCODE-AWG/tss-annotation/blob/master/long_read/pacbio_to_tss.py`) (`--window-size=50 --raw-counts --expression-threshold=0`), yielding a bed entry for each TSS consisting

of a wide peak, narrow peak, and a summit for each TSS. Resultant TSSs were filtered by requiring each one to be supported by at least 20 reads.

Identification of TESs from short-read data

Bam reads from the short-read Split-seq 464 cells were queried for those that contained a run of 20bp where at least 10 bases were "A"s using Cutadapt v. 2.10 (`-g AAAAAAAAAAAAAAAAAAAA -e 0.5 -O 20`), which represent reads that are more likely to contain a true 3' end[187]. TESs were called on the filtered bams using the ENCODE PacBio TSS caller (which despite its name works on short reads as well) `https://github.com/ENCODE-AWG/tss-annotation/blob/master/long_read/pacbio_to_tss.py`) (`--window-size=50 --raw-counts --expression-threshold=0 --tes`), yielding a bed entry for each TES consisting of a wide peak, narrow peak, and a summit for each TES. Resultant TESs were filtered by requiring each one to be supported by at least 20 reads.

Processing C2C12 CAGE data

CAGE data was downloaded from GEO accession GSE2158[163,188]. Wig files corresponding to CAGE data from days 0 and 9 of C2C12 differentiation were converted to bed format using bedops wig2bed[189] and lifted over from the mm9 genome to the mm10 genome using UCSC's liftOver tool (`-minMatch=0.95`)[190]. Resultant bed peaks were concatenated.

Processing C2C12 PolyA-seq data

PolyA-seq data was downloaded from GEO accession GSE62001[191,192]. Entries in the provided expression matrix were filtered for those belonging to the "C2C12.Pro" (proliferating C2C12) and "C2C12.Diff" (4-day differentiation C2C12) categories. The data was then converted into bed format using a custom script and lifted over from mm9 to mm10 using the UCSC liftOver tool (`-minMatch=0.95`)[190].

Intersecting TSSs with validation datasets

A combined TSS validation bed file was made using the proximal enhancer and promoter ENCODE cCREs[164], GENCODE vM21 TSSs[193], our snATAC-seq pseudobulk peaks, and CAGE peaks[163]. The filtered TSSs for both bulk (22,938) and LR-Split-seq (23,996) were intersected with the combination bed file using bedtools intersect 2.30.0 with default parameters, meaning minimum of 1bp overlap between the TSSs and the combined validation set (2,057,291 regions)[186].

Intersecting TESs with validation datasets

A combined TES validation bed file was made using our snATAC-seq pseudobulk peaks and polyA-seq peaks[192]. Similar to TSS validation, bedtools intersect with default overlap settings (1bp) was used to determine the number of overlaps between our filtered TESs for both bulk (14,120) and LR-Split-seq (12,521) and the combined validation set (205,853 regions).

snATAC-seq and TSS integration

TSS regions identified in LR-Split-seq in bed format were used to calculate activity at each TSS through the Signac interface. Normalized expression values and normalized TSS activity values were averaged across the three groups of cells (MB, $Pax7^{hi}$, and $Myog^{hi}$) and a pseudocount of 1 was added to each TSS. Fold change in expression and activity separately was calculated by dividing the TSS values of one group by another group, such as $Myog^{hi}$/MB. The log2 fold change for each TSS was then plotted for both expression (x-axis) and activity (y-axis), revealing TSSs with chromatin profiles and expression in agreement at the upper right and bottom left sectors. Twice the standard deviation of each dataset is indicated by black dashed lines. (Fig. 2.4f-2.4h). We determined whether a TSS was expressed and/or accessible using a read cutoff of 2 for LR-Split-seq data and a cutoff of 1000 for normalized snATAC-seq data (Fig. 2.12g).

LR-Split-seq TSS quantification and differential TSS testing

TSS expression was quantified from the LR-Split-seq data starting from the TALON read annotation file (Table S1), which tracks the start and end coordinates of every read. Read starts were converted to a read start bed file and expanded to include ±25 bp from the true start. Finally, the read start bed was intersected using bedtools with the filtered LR-Split-seq TSSs (Table S8), requiring at least 1 bp of overlap. The number of reads per TSS was then computed by counting all of the reads assigned to each TSS. Testing on the TSS level for the LR-Split-seq data was performed as in Joglekar et. al., 2021[133]. Tests were performed on the LR-Split-seq Leiden clusters for MB nuclei vs. *Myog*^{hi} nuclei (clusters LR1-LR2 vs. LR6-LR7) and for MB nuclei vs. *Pax7*^{hi} (clusters LR1-LR2 vs. LR4-LR5). Genes with significant TSS switching were required to have a corrected p val ≤ 0.05 and a change in percent isoform usage per condition of ≥ 10, and a minimum number of reads per gene per tested condition of 10. A custom UCSC Track Hub displaying pseudobulk snATAC peaks per cluster, LR-Split-seq reads used for TSS calling per cluster, ENCODE cCREs, and GENCODE vM21 transcript models can be found at `https://github.com/erebboah/c2c12_trackhub`.

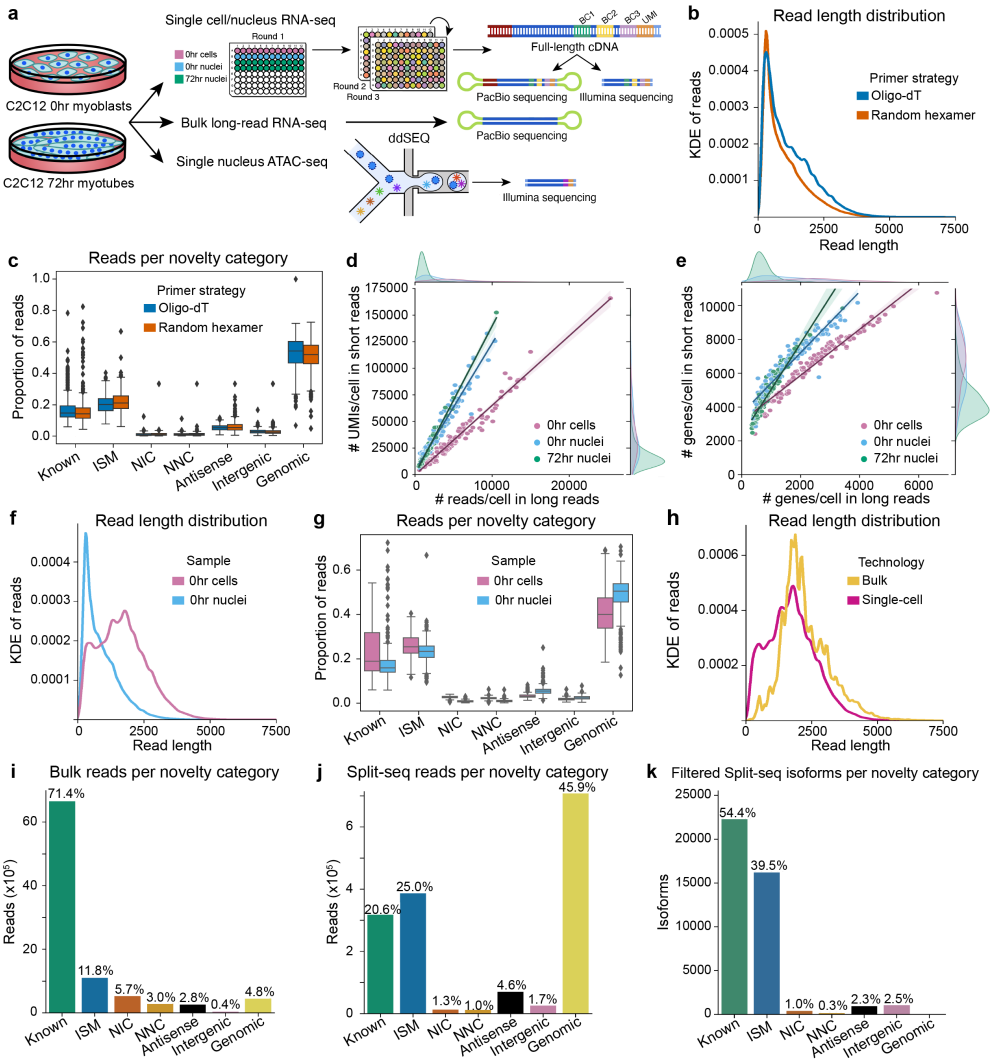

Figure 2.1: **Technical comparisons in LR-Split-seq and bulk long-read RNA-seq. a,** Schematic diagram of experimental design. Single cell/nucleus LR-Split-seq, short-read Split-seq, bulk long-read RNA-seq, and single nucleus ATAC-seq were performed on C2C12 0 h myoblasts and 72 h differentiating cells. The same single-cell/UMI-barcoded cDNA was used in both short-read and long-read sequencing. **b,** Kernel density estimation (KDE) of read length distribution of oligo-dT primed reads (blue) compared to random hexamer primed reads (orange). **c,** Proportion of oligo-dT/random hexamer reads in each cell for each novelty category. **d,** Comparison of number of reads and **e,** genes detected between short and long reads. Cells are labeled by sample type (0 h cells in pink [regression m = 1.4 and 6.5 in genes and reads respectively], 0 h nuclei in blue [regression m = 1.8 and 12.0 in genes and reads respectively], and 72 h nuclei in green [regression m = 2.7 and 13.9 respectively]) and marginals on the top and right indicate their distributions. **f,** KDE read length distribution of 0 h cells (pink) compared to 0 h nuclei (blue) reads, not including genomic reads. **g,** Proportion of 0 h cell (pink)/nuclei (blue) reads per cell/nucleus per novelty category. **h,** KDE read length distribution of bulk long reads (yellow) compared to single-cell long reads (magenta), not including genomic reads. **i,** Unfiltered reads per novelty category in bulk long-read data and **j,** LR-Split-seq data. **k,** Filtered isoforms per novelty category across all cells in LR-Split-seq data.

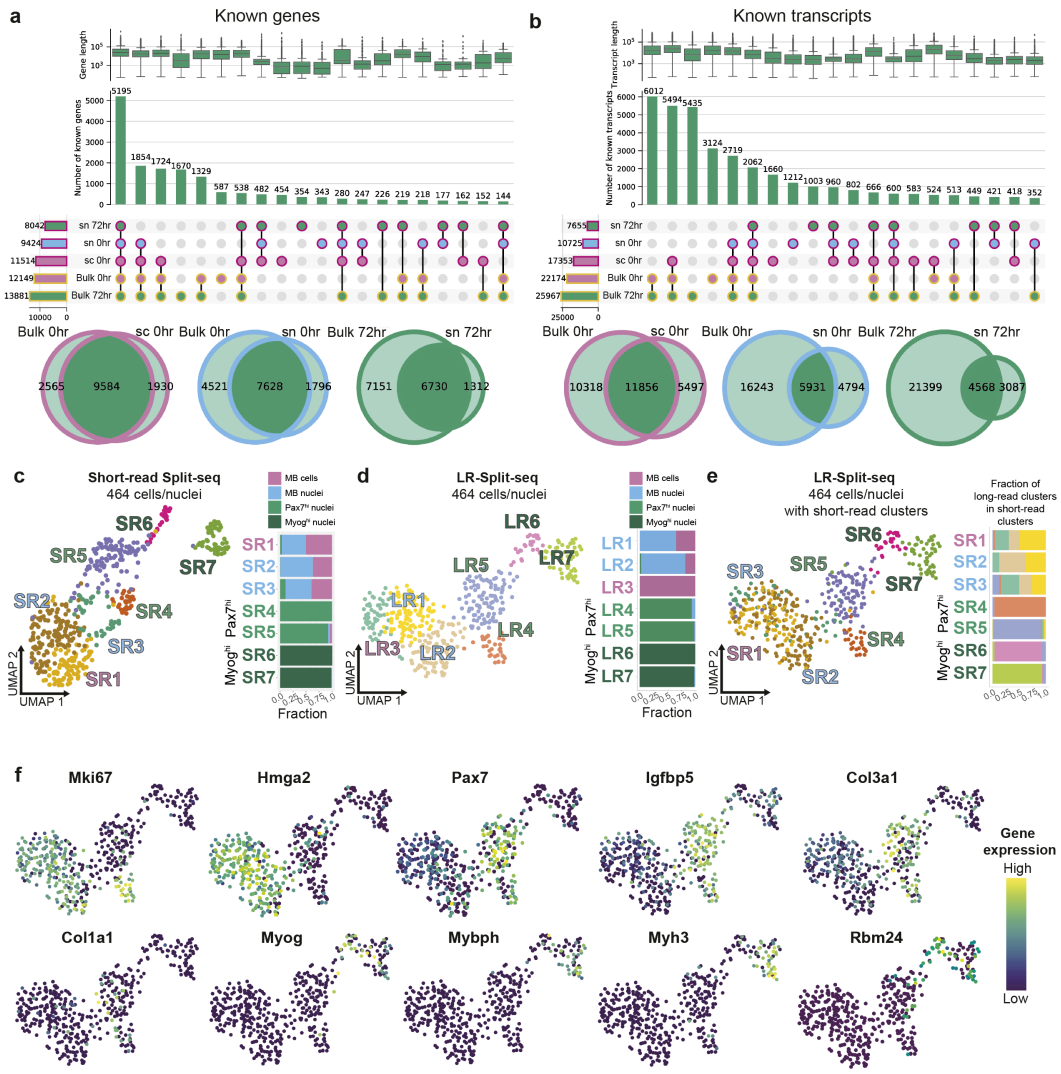

Figure 2.2: **LR-Split-seq in C2C12 0 h and 72 h samples recapitulates results from companion bulk and standard short-read Split-seq. a,** Upset plots of known genes found in bulk compared to LR-Split-seq data across all samples. Bars on the left indicate set size, circles indicate combinations of samples, and bars on top indicate the number of genes found in each combination (first 20 combinations shown). Outline colors indicate technology (bulk in yellow, single-cell or single-nucleus in magenta) and fill colors indicate sample type (72 h nuclei in green, 0 h nuclei in blue, and 0 h cells in pink for single-cell data; 72 h in green, 0 h in pink for bulk data). Box plots above indicate gene length distribution for each intersection. Venn diagrams below summarize the overlaps between bulk (left) and single-cell or single-nucleus (right), for each sample type. Sample type is indicated by outline color. **b,** Upset plot and Venn diagrams of known transcripts found in bulk data and LR-Split-seq data (first 20 combinations shown). **c,** UMAP of 464 short-read Split-seq cells/ nuclei labeled by 7 Leiden clusters (S) and breakdown of cell type per cluster: 110 0 h cells (pink), 145 0 h nuclei (blue), and 209 72 h nuclei (*Pax7*^{hi} in green and *Myog*^{hi} in dark green). **d,** UMAP of 464 LR-Split-seq cells/nuclei using gene-level data labeled by 7 Leiden clusters (LR) and **e,** Leiden cluster ID of matching short-read data (SR) shown in c, as well as long-read cluster makeup of each short-read cluster. **f,** Expression of marker genes, dark blue = lowly expressed, yellow = highly expressed.

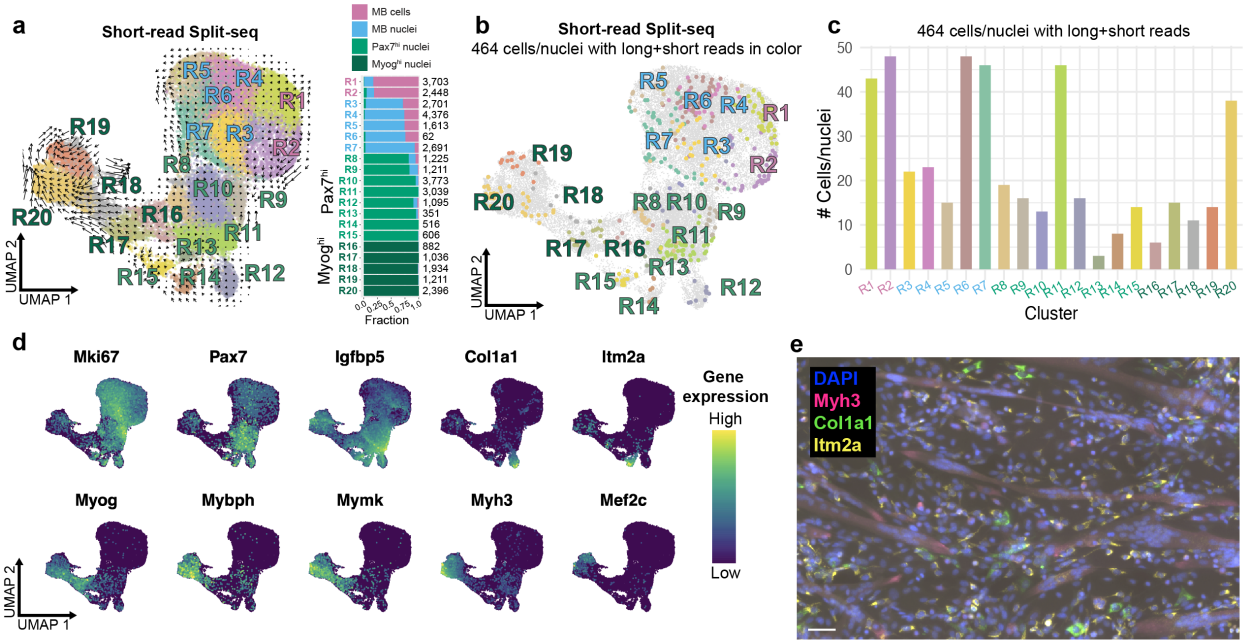

Figure 2.3: **Short-read Split-seq analysis.** **a,** UMAP of 36,869 short-read Split-seq cells/nuclei labeled by 20 Leiden clusters (R) with RNA velocity field trajectories and breakdown of cell type per cluster with number of cells per cluster: 7797 0 h myoblast cells (pink), 10,194 0 h myoblast nuclei (blue), 18,878 72 h nuclei (*Pax7*^{hi} in green and *Myog*^{hi} in dark green). **b,** UMAP of short-read Split-seq cells/nuclei with the 464 cells with matching long reads in color corresponding to R1-R20. **c,** Histogram of the number of the 464 cells/nuclei per R1-R20. **d,** Distribution of expression of marker genes; dark blue = lowly expressed, yellow = highly expressed. **e,** Visualization of transcripts in mononucleated cells and myotubes at the 72 h differentiation timepoint. Blue = DAPI, pink = *Myh3*, green = *Col1a1*, yellow = *Itm2a*. Scale bar: 50 µm.

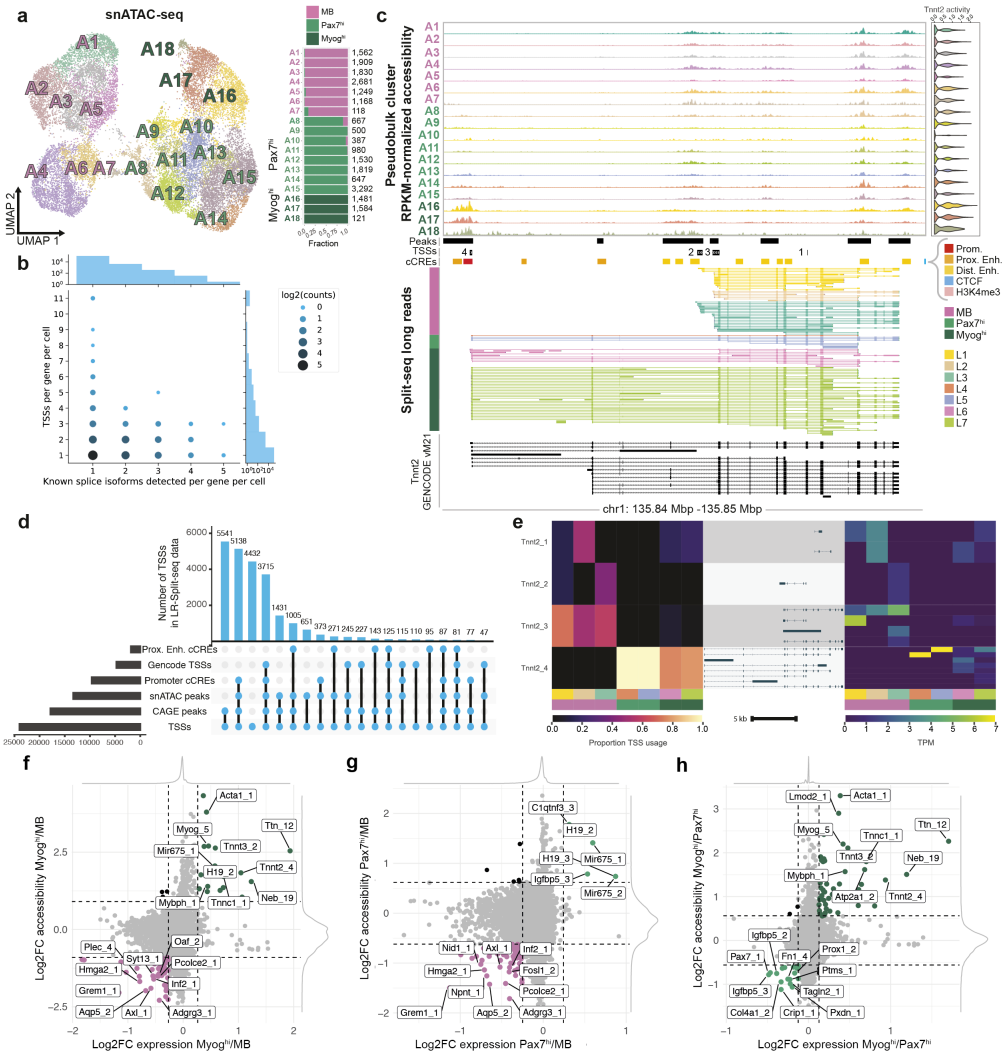

Figure 2.4: **Identification of TSSs from LR-Split-seq and integration with snATAC-seq. a,** UMAP of 23,525 snATAC- seq nuclei labeled by 18 Leiden clusters (A) and breakdown of cell type per cluster with number of cells per cluster on right: 10,508 0 h myoblast nuclei (pink) and 13,017 72 h nuclei (*Pax7*^{hi} in green and *Myog*^{hi} in dark green). **b,** Bubble plot of the number of distinct known splice isoforms per gene per cell compared to the number of distinct TSSs per gene per cell in LR-Split-seq. **c,** Track plot of alternative *Tnnt2* TSS usage between 72 h differentiating cells and 0 h myoblasts. From top to bottom: clustered snATAC-seq pseudobulk peaks, merged psuedobulk peaks, TSS regions called from LR-Split-seq, ENCODE cCREs, clustered LR-Split-seq reads used to call TSSs, and comprehensive set of GENCODE vM21. **d,** Validation of TSSs found in LR-Split-seq using four external datasets and snATAC-seq pseudobulk peaks (first 20 intersections shown). **e,** Left, proportion of TSS-assigned reads in LR-Split-seq clusters from each identified *Tnnt2* TSSs. Right, expression of each TALON filtered *Tnnt2* isoform in LR-Split-seq clusters with corresponding transcript models associated with each *Tnnt2* TSS. **f,** Comparison of log2 fold change (LFC) in expression and accessibility across identified TSSs: *Myog*^{hi} (+LFC) compared to MB (-LFC), **g,** *Pax7*^{hi} (+LFC) compared to MB (-LFC), **h,** *Myog*^{hi} (+LFC) compared to *Pax7*^{hi} (LFC).

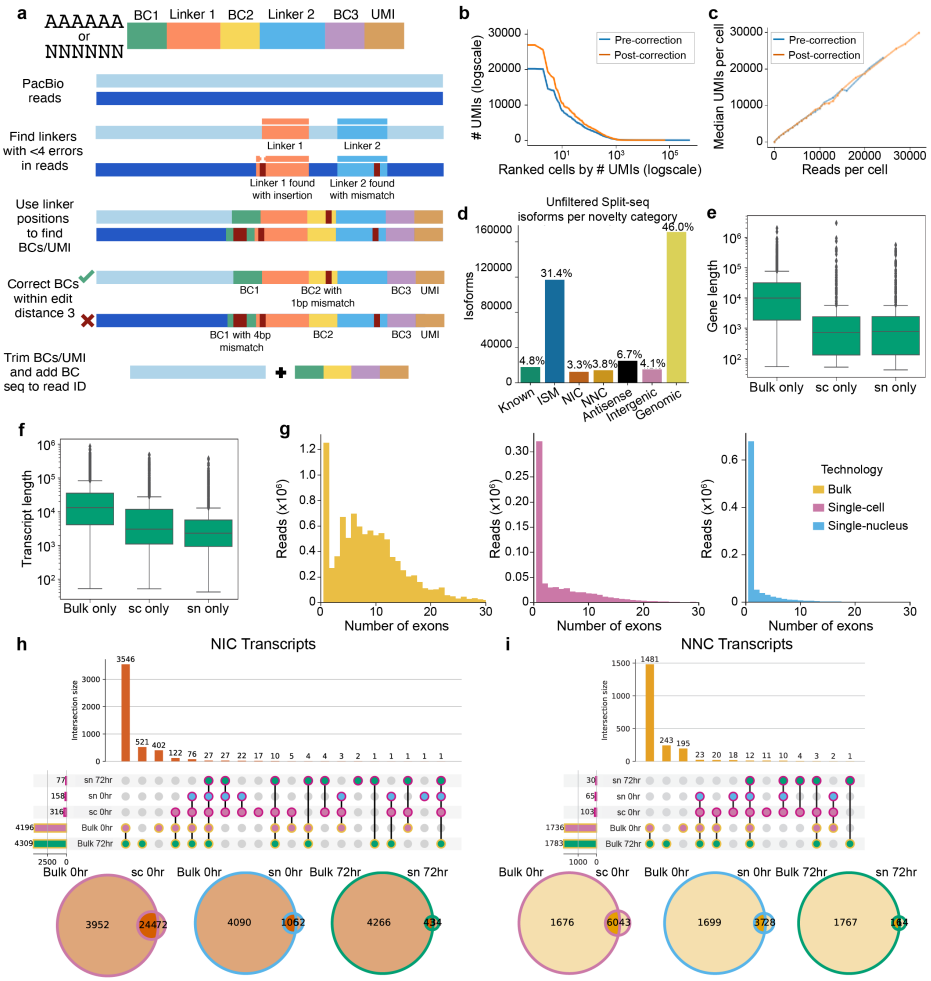

Figure 2.5: **LR-Split-seq preprocessing, QC, and additional analysis. a,** Schematic diagram of LR-Split-seq demultiplexing strategy. **b,** UMI per ranked barcode plots before and after barcode correction (both axes log scaled). **c,** Median number of UMIs per cell binned by reads per cell before and after barcode correction. **d,** Unfiltered isoforms per novelty category across all cells in LR-Split-seq data. **e,** Gene lengths of annotated genes detected in bulk only, single-cell only, and single-nucleus only (log scale). **f,** Transcript lengths of annotated transcripts detected in bulk only, single-cell only, and single-nucleus only (log scale). **g,** Distribution of number of exons in bulk long reads (yellow), single-cell long reads (pink), and single-nucleus long reads (blue). **h,** Upset plot of novel in catalog (NIC) transcripts that passed filtering found in bulk data compared to single cell data across all samples. Bars on the left indicate set size, circles indicate various combinations of samples, and bars on top indicate the number of genes found in each combination. Outline colors indicate technology (bulk in yellow, single-cell in magenta) and fill colors indicate sample type (72 h nuclei in green, 0 h nuclei in blue, and 0 h cells in pink for single-cell data; 72 h in green, 0 h in pink for bulk data). Box plots above indicate gene length distribution for each intersection. Venn diagrams below summarize the overlaps between bulk (left) and single-cell or singlenucleus (right), for each sample type. Sample type is indicated by outline color. **i,** Upset plot and Venn diagrams of novel not in catalog (NNC) transcripts that passed filtering found in bulk data and single-cell data.

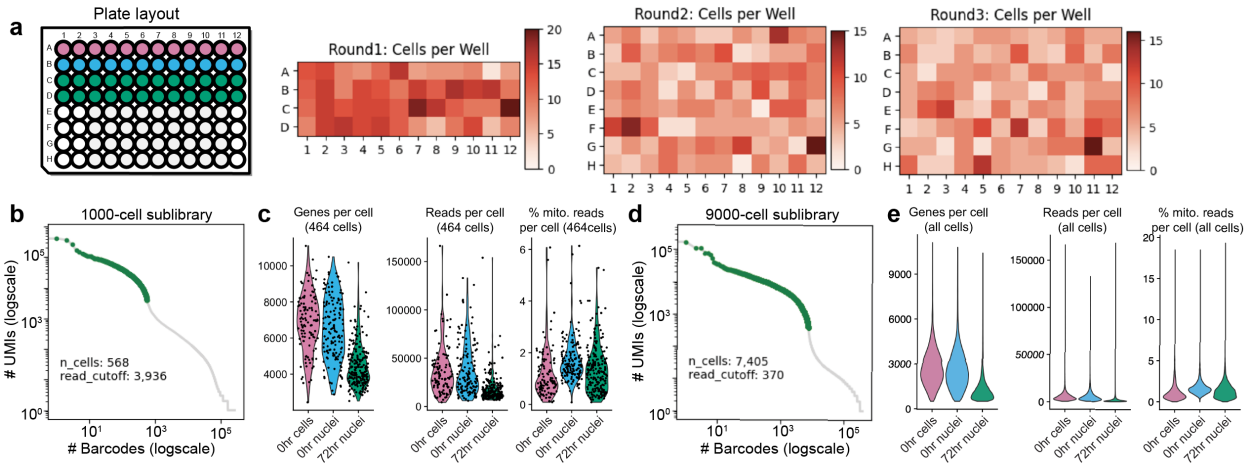

Figure 2.6: **Short-read Split-seq QC a,** Schematic of sample type per well in the first round of barcoding (pink = 0 h cells, blue = 0 h nuclei, and green = 72 h nuclei). Panels to the right show the number of cells per well across each round of barcoding for a 9000-cell sublibrary. **b,** UMI per cell knee plots for the 1000-cell sublibrary sequenced with both long and short reads indicating a threshold of 3,936 reads per cell, leaving 568 cells before additional QC. **c,** Violin plots of scRNA-seq QC metrics after filtering for the 464 cells only. **d,** An example knee plot for a 9000-cell sublibrary indicating a threshold of 370 reads per cell, leaving 7,405 cells before additional QC. **e,** Violin plots of scRNAseq QC metrics after filtering for all cells.

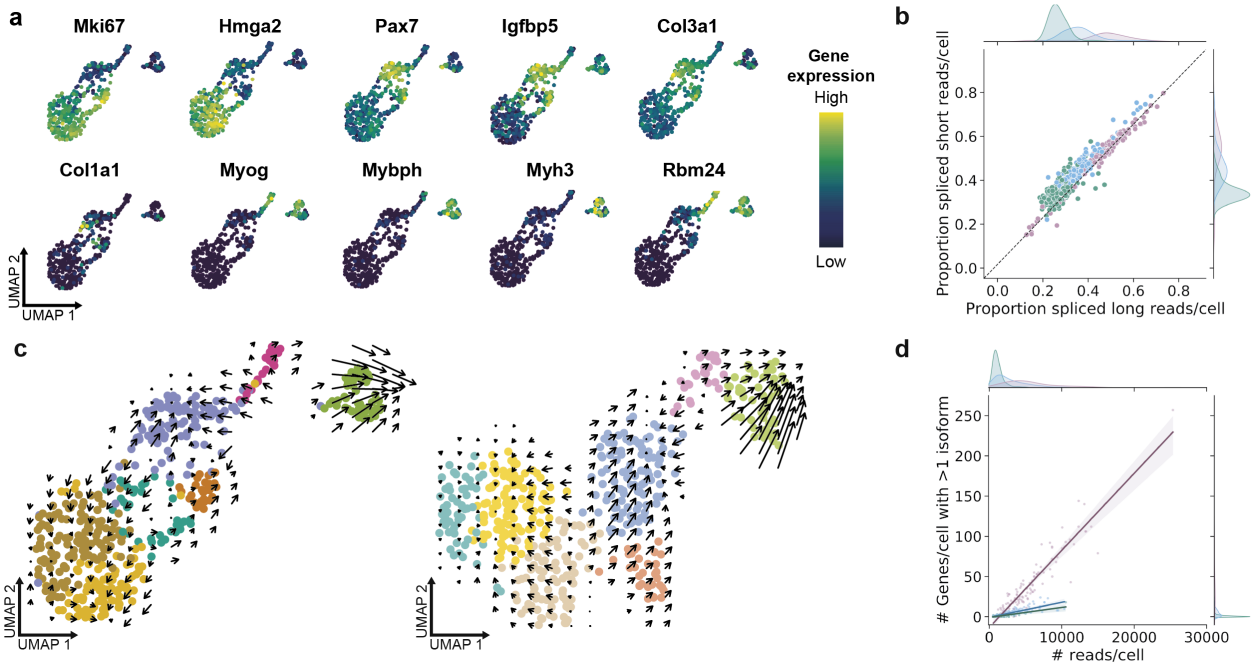

Figure 2.7: **Short-read and LR-Split-seq additional analysis. a,** Distribution of marker genes within the 464-cell UMAP (dark blue = lowly expressed, yellow = highly expressed). **b,** Proportion of spliced vs. unspliced reads per cell in short-read Split-seq and LR-Split-seq from RNA velocity analysis. Cells are labeled by sample type (0 h cells in pink, 0 h nuclei in blue, and 72 h nuclei in green) and marginals on the top and right indicate their distributions. **c,** Short-read (left) and LR-Split-seq (right) UMAPs for 464 cells with RNA velocity field trajectories indicated by arrows. **d,** Isoform complexity (Number of genes with more than one isoform per cell) vs. number of reads per cell, colored by sample type (0 h cells in pink, 0 h nuclei in blue, and 72 h nuclei in green).

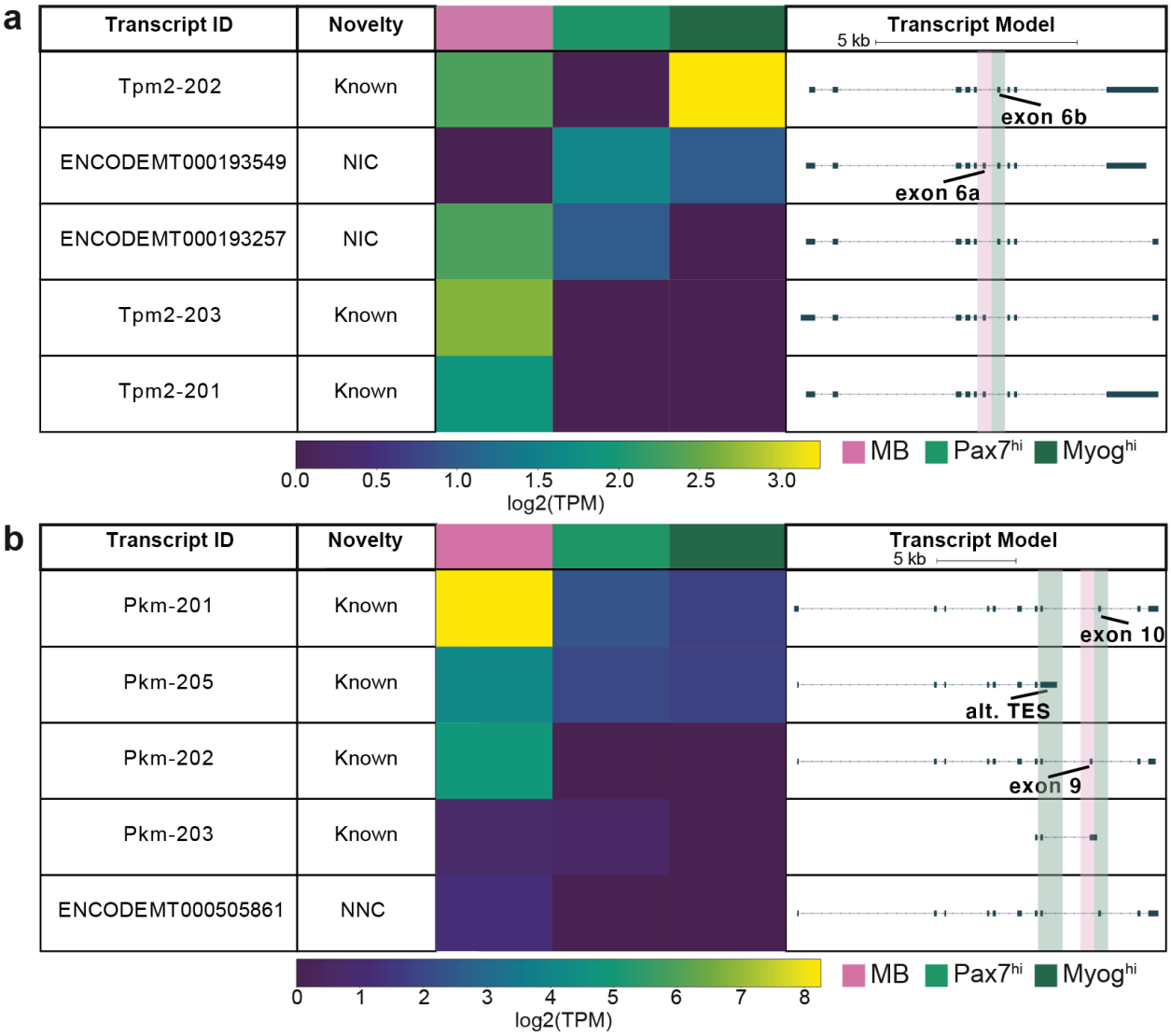

Figure 2.8: **Swan analysis of *Tpm2* and *Pkm* isoforms. a,** Gene report made by Swan for *Tpm2*. Relative expression of each isoform, separated by 0 h MB cells, 72 h *Pax7*^{hi} nuclei, and 72 h *Myog*^{hi} nuclei plotted alongside the isoform's name, transcript novelty, and structure. Exons 6a and 6b, known to be alternatively spliced during C2C12 differentiation, are highlighted. **b,** Gene report made by Swan for *Pkm*, separated by the same cell types. Mutually exclusive exons 9 and 10 as well as alternative TES in Pkm-205 are highlighted.

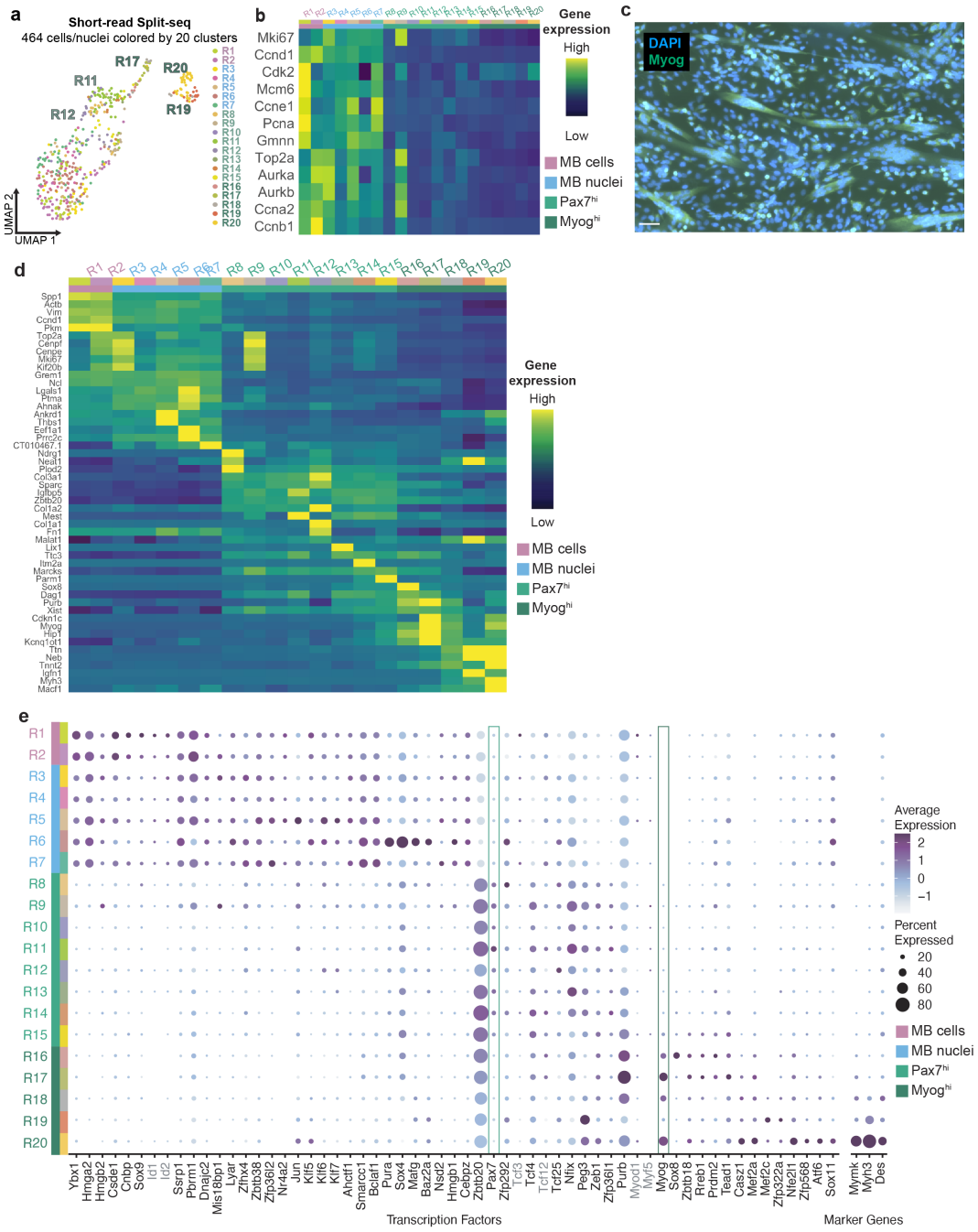

Figure 2.9: **Additional analysis of 38,000-cell short-read Split-seq data. a,** UMAP of 464 cells with both short and long reads colored by 20 clusters derived using 36,869 short-read cells. **b,** Heatmap of cell cycle marker genes in the 20 clusters. **c,** Visualization of *Myog* in mononucleated cells and myotubes at the 72 h differentiation timepoint. Blue = DAPI, green = *Myog*. Scale bar: 50 µm. **d,** Heatmap of marker genes in the 20 clusters (dark blue = low expression, yellow = high expression). **e,** Dot plot of transcription factors and marker genes involved in myogenesis found from differential expression testing and/or literature. Genes that did not pass the differential expression threshold yet are of interest in the system and significantly expressed in prior classic bulk data are colored grey (*Id1*, *Id2*, *Myod1*, *Myf5*, *Tcf3*, and *Tcf12*).

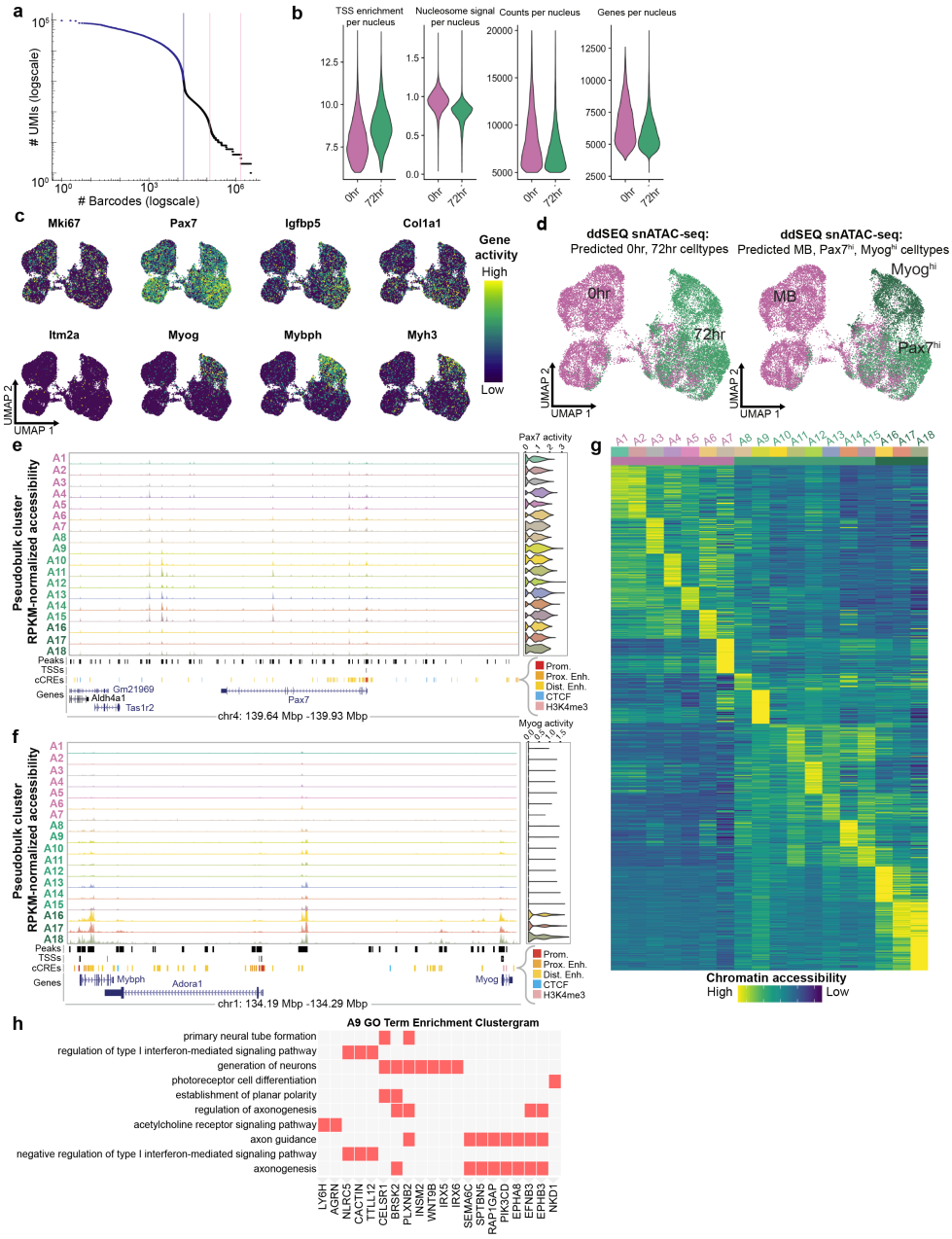

Figure 2.10: **Additional analysis/QC of snATAC-seq. a,** UMI per barcode knee plot for an example snATAC-seq library (0 h, 6,782 nuclei). **b,** Violin plots of snATAC-seq QC metrics after filtering >6 TSS enrichment, <20,000 reads, and >5,000 reads per nucleus. **c,** Distribution of marker genes within the UMAP colored by gene activity score (dark blue = low activity, yellow = high activity). **d,** Integration of scRNA-seq and snATAC-seq data, labeled by cell type (0 h in pink and 72 h in green on left; MB in pink, *Myog*^{hi} in dark green, and *Pax7*^{hi} in light green on right). **e,** Pseudobulk peaks per cluster spanning the *Pax7* locus. TSS track indicates TSSs called from LR-Split-seq data. **f,** Pseudobulk peaks spanning the *Myog* and *Mybph* loci. **g,** Heatmap of top 50 marker regions in the 18 snATAC-seq clusters (dark blue = low accessibility, yellow = high accessibility). **h,** Cluster A9 GO term enrichment clustergram. Examples of genes associated with A9 marker peaks belonging to the GO terms in rows are indicated in red.

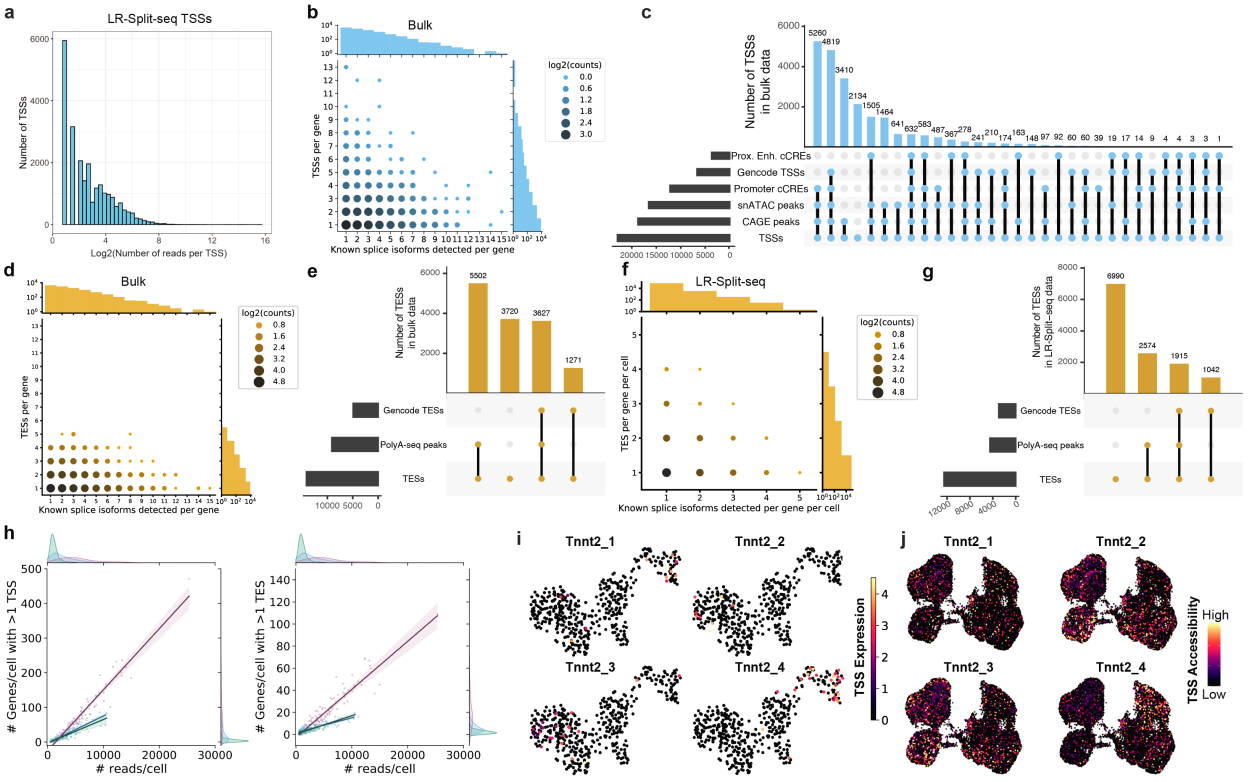

Figure 2.11: **Identification and validation of TSSs/TESs from long-read data. a,** Histogram of number of LR-Split-seq reads supporting each TSS. **b,** Bubble plot of the number of distinct exon combinations (splice isoforms) detected per gene compared to the number of distinct TSSs detected per gene in bulk data. **c,** Validation of TSSs found in bulk long-read data using 4 external datasets (ENCODE proximal enhancer and promoter cCREs, GENCODE TSSs, and CAGE peaks) and our snATAC-seq pseudobulk peaks. **d,** Bubble plot of the number of distinct exon combinations (splice isoforms) detected per gene compared to the number of distinct TESs detected per gene found in long-read bulk data. **e,** Validation of TESs found in bulk long-reads using GENCODE TESs and polyA-seq data. **f,** Bubble plot of splice isoforms per gene per cell compared to TESs detected per gene per cell found in LR-Split-seq. **g,** Validation of TESs found in LR-Split-seq. **h,** TSS and TES complexity (Number of genes with more than one TSS / TES per cell) vs. number of reads per cell, colored by sample type (0 h cells in pink, 0 h nuclei in blue, and 72 h nuclei in green). **i,** LR-Split-seq TSS expression for the 4 identified *Tnnt2* TSSs. **j,** snATAC accessibility for the 4 identified *Tnnt2* TSSs.

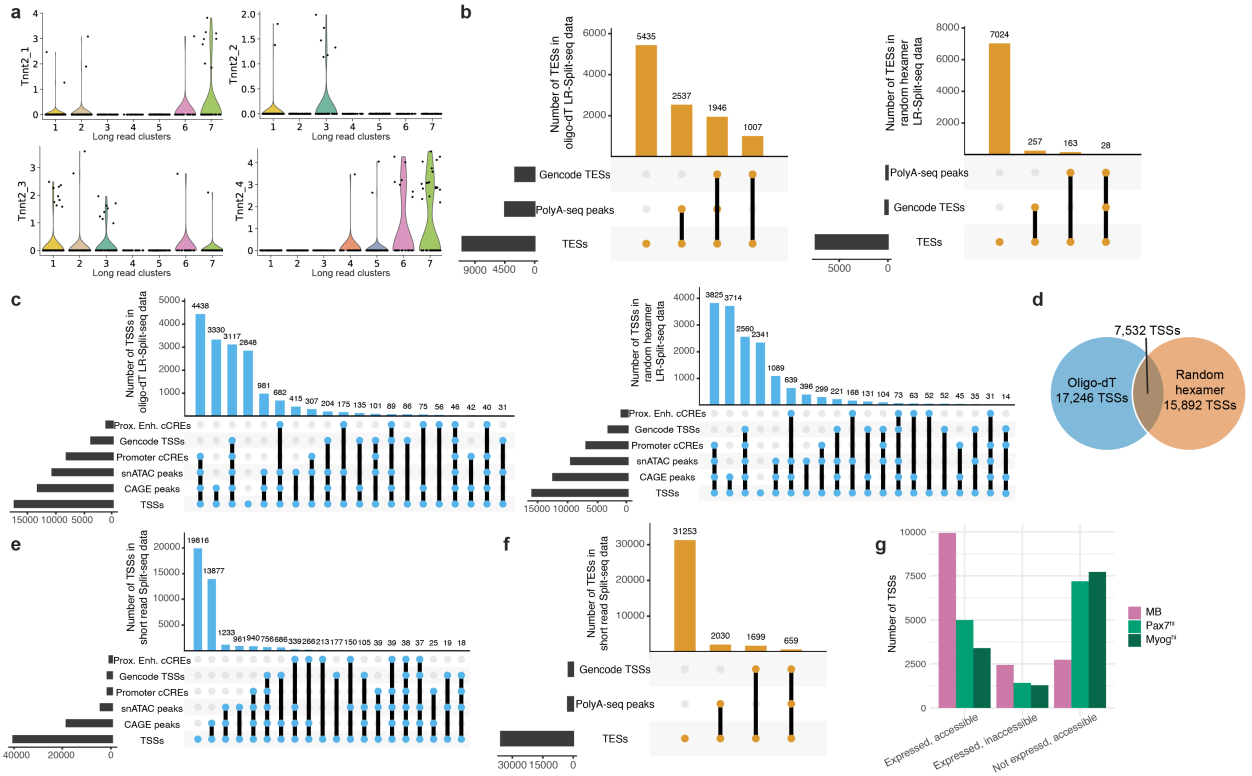

Figure 2.12: **Identification and validation of TSSs/TESs from oligo-dT and random hexamer long reads, and short-read data. a,** Violin plots of TSS expression per long read cluster for the 4 identified *Tnnt2* TSSs. **b,** Validation of TESs found in LR-Split-seq data split by oligo-dT primed reads (left) and random hexamer primed reads (right). **c,** Validation of TSSs found in LR-Split-seq data split by oligo-dT primed reads (left) and random hexamer primed reads (right). **d,** Venn diagram of oligo-dT and random hexamer TSSs with 7,532 TSSs overlapping by at least 1 bp. **e,** Validation of TSSs found in short-read Split-seq data using reads fully containing the TSO (template switching oligo). **f,** Validation of TESs found in short-read Split-seq data using reads containing a 20bp poly-A sequence allowing for 50% mismatches. **g,** Comparison of thresholded expression and accessibility of TSSs split by cell type.

Chapter 3

The ENCODE mouse postnatal developmental time course identifies regulatory programs of cell types and cell states

3.1 Abstract

Postnatal genomic regulation significantly influences tissue and organ maturation but is under-studied relative to existing genomic catalogs of adult tissues or prenatal development in mouse. The ENCODE4 consortium generated the first comprehensive single-nucleus resource of postnatal regulatory events across a diverse set of mouse tissues. The collection spans seven postnatal time points, mirroring human development from childhood to adulthood, and encompasses five core tissues. We identified 30 cell types, further subdivided into 69 subtypes and cell states across adrenal gland, left cerebral cortex, hippocampus, heart,

and gastrocnemius muscle. Our annotations cover a span of both known and novel dynamics ranging from early hippocampal neurogenesis and a new sex-specific adrenal gland population during puberty. We used robust Latent Dirichlet Allocation with a curated vocabulary of 2,701 regulatory genes to identify regulatory topics linked to cell type differentiation, subtype specialization, and transitions between cell states. Shared topics emerged in cycling cells of the adrenal gland and heart, tissue-resident macrophages, neural cell types, and endothelial cells across multiple tissues. Cell-type-specific topics are enriched in transcription factors and microRNA host genes, while chromatin regulators dominate mitosis topics. Corresponding chromatin accessibility data reveal dynamic and sex-specific regulatory elements, with enriched motifs matching transcription factors in regulatory topics. Together, these analyses provide insights into postnatal development across various tissues through the lens of the factors regulating transcription.

3.2 Introduction

The postnatal period is a critical phase in an individual's life marked by pivotal processes such as physical and cognitive development, social and emotional interactions, as well as sensory and metabolic maturation. Both humans and mice undergo significant changes during the postnatal period, including puberty, with sex-specific growth and maturation of their bodies, as well as the advancement of motor skills such as crawling, walking, jumping, and running. Cell type specialization and cell state transitions underlie these biological processes[107,121]. Cell types maintain a stable, heritable identity, defined by shared characteristics such as molecular markers, morphology, location, and functional properties[109,110]. In contrast, cell states represent dynamic variations within a cell type, responding to environmental cues, developmental stages, or physiological changes. These variations involve shifts in gene or protein expression and epigenetic modifications without altering the fundamental

cell type[109,110]. Cell types establish cellular identity, while cell states contribute to diversity and plasticity within a specific cell type[109,110]. For example, postnatal growth of skeletal muscle occurs through the overall expansion of myofibers as well as the proliferation of myonuclei within them, leading to the emergence of distinct type I and type 2 skeletal myofibers with specific contractile properties[112,113,118,121]. While the myonuclei within muscle cells are defined as a stable cell type, exercise training can induce cell state transitions between type 1 and type 2 fibers[121]. To fully describe the specialization of cell types and transitions between cell states, comprehensive characterization of molecular intermediates such as gene expression and chromatin accessibility must be performed at the single-cell level.

Existing single-cell and single-nucleus catalogs primarily capture limited timepoints, focusing on either prenatal development or aging adults. The Tabula Muris Consortium, a widely used resource, recently captured over 350,000 cells in 6 age groups and 23 tissues and organs[194], building on their previous *Tabula Muris* catalog of 100,000 cells from 20 organs and tissues using single-cell RNA-seq (scRNA-seq)[195]. The *Tabula Muris Senis* focused on 1- to 30-month-old mice and identified 155 cell types, averaging around 800 cells per tissue[194]. Comparative analysis of gene expression across cell types from 3, 18, and 24-month-old mice suggested that certain cell types such as microglia exhibit an intermediate cell state before transitioning to an aged transcriptional profile[194]. In a focused approach, the systematic dissection of regions in the adult mouse cortex and hippocampus of the Allen Brain Atlas followed by scRNA-seq of 1.3 million cells has produced a comprehensive cell type taxonomy that aligns with the spatial arrangement of the brain[196]. Although 42 unique subclasses of predominantly GABAergic and glutamatergic neurons were identified, the annotation lacks expected mouse adult stem cells in the brain such as oligodendrocyte precursor cells and neuronal progenitor cells. To provide insights into mouse prenatal development, the ENCODE3 mouse embryo project profiled 12 whole tissues from embryonic day 10.5 to birth using bulk RNA-seq, as well as at the single-nucleus level in forelimb[84]. This prenatal single-nucleus timecourse of 91,557 total nuclei and 25 cell types revealed dynamic changes in cell type

composition and emergence of multiple lineages during skeletal myogenesis in the mouse forelimb. In contrast, our snRNA-seq study spans five core tissues from just after birth to late adulthood at comparable depth to the forelimb time course, pinpointing 99 distinct cell types and states. Our dataset includes an average of around 87,000 nuclei per tissue across 7 timepoints, incorporating 10x Multiome nuclei at two key timepoints.

An ongoing challenge in single-cell RNA-seq analysis is to identify and associate groups of genes with meaningful traits. When traits such as sex and age are defined in the data, differential expression analysis facilitates the direct comparison of genes enriched in one group compared to another. However, single-cell RNA sequencing notoriously reveals novel cell types and states without clear definitions. In such cases, identifying genes associated with these populations presents a significant challenge. While co-expression network analysis is a popular approach for grouping genes into modules without predefined annotations[122,197?] , it limits each gene to a single module. Another favored method applies Latent Dirichlet Allocation (LDA), also known as topics modeling, to gene expression data. LDA was originally introduced for population genetics[125], then a few years later in natural language processing using machine learning[124]. In the context of written documents, LDA is a generative model that groups words into topics, allowing multiple topics to be associated with a single document, and assigns a numeric weight to each word in every topic. A word may have a high weight in more than one topic, or in none. More recently, LDA has been repurposed for single-cell RNA-seq to model gene expression by considering genes as words, cells as documents, and latent biological processes as topics[126,127]. The mixed membership flexibility of LDA aligns with biological reality, where a gene may be repurposed in multiple cellular programs. Analyzing gene weights between topics facilitates the comparison of more ambiguous traits associated with topics, such as dynamic cell types and states, in addition to age and sex.

The core ENCODE4 mouse time course captures postnatal development at key timepoints

across cerebral cortex, hippocampus, heart, skeletal muscle, and adrenal glands, encompassing 436,440 total nuclei. We apply robust LDA (rLDA) using Topyfic with a curated vocabulary of 2,701 regulatory mouse genes[128]. We recover 82 topics associated with 45 cell types and states including adult stem cells, tissue-resident macrophages, and general proliferation. Using this specific vocabulary allows us to capture cellular programs controlled by transcription factors (TFs) as well as other transcriptional and chromatin regulators such as coactivators, microRNAs, and histone modifiers, and compare them across diverse tissues. Finally, corresponding chromatin accessibility from 10x Multiome at two timepoints ties TFs within our regulatory topics to age-specific and sex-specific cell type- and state-specific regulatory element activity.

3.3 Results

The ENCODE4 mouse single-nucleus RNA dataset

For the final phase of the ENCODE Consortium, we comprehensively map the mouse polyadenylated RNA transcriptome at the single-nucleus level across 5 coordinated tissues at 7 timepoints in B6/CAST F1 hybrid mice, spanning from postnatal day (PND) 4 to late adulthood (18-20 months) using the Parse Biosciences combinatorial barcoding platform[88,89] (Fig. 3.1a). Complementary genome-wide datasets, including bulk short-read RNA-seq, long-read RNA-seq, microRNA-seq, and chromatin accessibility are also available for matching samples at some or all timepoints (Fig. 3.1b). Both polyadenylated RNA and chromatin accessibility were measured in the same single nuclei across all five tissues at PND 14 and 2-month timepoints using the 10x Multiome platform[87]. Notably, this mouse time course mirrors the majority of the human postnatal lifespan, capturing crucial developmental stages including the dynamic period of puberty and other key milestones in the transition

from infancy to adulthood.

We recovered 83,467 adrenal gland nuclei, 112,118 left cerebral cortex nuclei, 78,168 hippocampus nuclei, 92,808 heart nuclei, and 69,879 skeletal muscle nuclei, collectively expressing 47,707 genes (including protein coding, pseudogene, lncRNA, or microRNA gene biotypes). We annotated each tissue separately for a combined total of 188 clusters, 69 subtypes and states, and 30 major cell types (Fig. 3.5, 3.6, 3.7, 3.8, 3.9, Methods). Tissues were clustered with similar resolutions, and each cluster was annotated using established marker genes, expert consultations, cluster marker gene identification, literature review, and label transfer from reference datasets where applicable[196,198–200](Methods). Annotation occurred across three hierarchical levels: "subtypes", "cell types", and "general cell types". Every cluster was assigned a single subtype, with larger subtypes comprising multiple clusters. During the annotation process, cell states were tracked within subtypes. This resolution encompasses specialized myonuclei located beneath the neuromuscular junction, as well as dynamic sex-specific layers within the adrenal cortex. In instances where two clusters exhibited identical marker gene expression, they were annotated similarly. This occurred for large and relatively homogenous cell types, such as vascular endothelial cells. Evaluation of the number of unique molecular identifier (UMI) counts and genes across cell types reveals reproducible patterns across tissues. Neural cell types such as neurons and adrenal medulla chromaffin cells consistently have more UMIs, and therefore a larger number of detected genes, compared to other cell types such as endothelial and immune cells regardless of the total number of nuclei within each respective cell type (Fig. 3.1c). The observed variation in the number of detected genes across cell types could reflect differences in underlying biological processes. Neural cell types may express a more extensive and dynamic array of transcriptional programs compared to other cell types, resulting in a higher number of actively transcribed genes within the nucleus.

Sex specific layers expand in the adrenal zona fasciculata during puberty before shrinking in late adulthood

Previous studies in B6J mouse adrenal gland characterized the X-zone, a mouse-specific cortical layer situated between the central medulla and the encasing zona fasciculata (ZF) in both male and female mice[107]. The mouse X-zone and the human fetal zone are both transient cortical layers originating from the fetal stage of development[107,201]. The human fetal zone disappears rapidly after birth, along with a decrease in steroid secretion, but is functionally similar to the human-specific zona reticularis in adults[201]. The mouse X-zone becomes detectable by PND 8 and fully emerges as a distinguishable layer by PND 14[107]. In female mice, this layer persists for several weeks during puberty until beginning to regress by PND 32 at the earliest, continuing regression during adulthood. During the first pregnancy, the entire X-zone disappears, while in non-pregnant mice, it undergoes gradual regression before disappearing between 3 and 7 months[107]. In male mice, the X-zone recedes entirely before PND 40[107]. While the human zona reticularis continues to produce androgens at lower levels after birth, increasing during puberty, mice adrenals lack expression of *Cyp17a1* and thus do not secrete androgens[202]. Instead, the X-zone is characterized by the expression of 20-alpha-hydroxysteroid dehydrogenase (*Akr1c18*), which has been shown to be induced by estrogen and downregulated by testosterone[107]. Additionally, *Pik3c2g*, a phosphoinositide 3-kinase involved in cell proliferation, survival, and metabolism is an X-zone marker[107]. Furthermore, thyroid nuclear hormone receptor beta1 (*Thrb*) shares X-zone-specific expression with *Akr1c18*. Despite the specificity of these markers, corresponding knockout mouse models lack any X-zone phenotype[107]. Sex-related factors and other molecules involved in the formation, maintenance, and regression of the X-zone reportedly have no specific expression in the X-zone. Thus, the function of the X-zone remains unclear despite the steroidogenic activity of the fetal adrenal cortex from which it originates.

We identify in males the X-zone counterpart, a large cluster of 8,104 male-specific ZF nuclei

that emerges from PND 25 to PND 36 and also regresses in later adulthood (Fig.3.1d, Fig. 3.5). Male nuclei make up 95% of the clusters we annotate as male-only ZF, while female nuclei make up 86% of X-zone clusters (4,505 nuclei). We find 303 differentially expressed genes with adjusted p-value < 0.01 and log2 fold change (LFC) > 1 upregulated in females compared to males in the X-zone and male-specific ZF, including *Xist* and *Tsix* as well as X-zone marker *Pik3c2g* (Methods). *Akr1c18* is not significantly upregulated, but still displays X-zone specific expression (Fig. 3.5). Ten of the genes upregulated in females are TFs, including *Thrb*, *Runx2*, *Irf8*, and *Nr3c1*. In males compared to females within sex-specific clusters, 666 genes are differentially expressed with adjusted p-value < 0.01 and LFC > 1, including Y-chromosome linked *Uty* and 35 TFs including *Esrrg* and *Hhex*. Considering these characteristics such as nucleus count, sex specificity, differentially expressed genes, and dynamics mirroring the X-zone in females, we designated the male ZF as a distinct subtype within the broader zona fasciculata in males and females.

Postnatal neurogenesis and glial maturation in the brain

The hippocampal dentate gyrus (DG) is one of the few brain regions that exhibits postnatal neurogenesis across several mammalian species, controversially including humans[108,203,204]. In mice and rats, the initial month of postnatal development marks a crucial transitional phase. The neurogenic processes and maturation trajectory of the granule cell population in early postnatal development align with those observed in adult neurogenesis[108]. The most significant maturation shift in the granule cell population occurs between PND 7 and 14[108]. During this period, neuronal progenitor cells (NPCs) expressing doublecortin (*Dcx*) become localized to the innermost region of the granule cell layer, signifying the establishment of the subgranular zone[108]. Adult neurogenesis occurs in this specialized niche, from which NPCs eventually migrate to the overlying granule cell layer and become integrated in hippocampal circuitry[203]. These conclusions are supported by our observation that 73% of early PND

10 and PND 14 DG nuclei belong to separate clusters than 92% of PND 25 and later DG nuclei. Pseudotime ordering from a starting node of cycling nuclei is consistent with real time, distinguishing PND 10 and PND 14 from later timepoints (Methods). Our findings suggest that in later timepoints, the predominant DG cell population is composed of mature *Calb1+* granule cells; however, approximately a quarter of all our immature *Dcx+* early DG cells persist into late adulthood (Fig. 3.1, 3.7). Their presence may have implications for cognitive functions such as synaptic plasticity, learning, memory formation, and stress resilience[205,206].

Glial maturation is also captured in both the hippocampus and cerebral cortex as a differentiation trajectory from oligodendrocyte precursor cells (OPCs) made up of predominantly early timepoints, though they are present throughout adulthood at lower proportions, to myelin-forming oligodendrocytes (MFOL), to mature oligodendrocytes (MOL) (Fig. 3.1d, 3.6, 3.7). Characterized by the expression of proteoglycan neuron-glial antigen *Cspg4*[207], homeodomain transcription factor *Nkx2-2*[207], and mitogen *Pdgfra*[208], OPCs constitute a highly dynamic and proliferative group of progenitors (Fig. 3.6, 3.7). The differentiation of OPCs into oligodendrocytes facilitates ongoing oligodendrocyte generation in adulthood, contributing to adaptive myelination and the capacity to regenerate myelin in response to injury or disease[208].

Cycling and perinatal populations in early postnatal stages of cardiac and skeletal myonuclei

Significant postnatal development occurs in both cardiac and skeletal muscle. In heart, growth is categorized into three phases after birth: hyperplasia until PND 4, rapid hypertrophy between PND 5 and 15, and slow hypertrophy from PND 15 onward[209]. In our data, proliferating cardiomyocytes marked by expression of *Top2a* and *Mki67* diminish by

PND 10, indicating that the first wave of growth is mainly due to cellular division (Fig. 3.1d). Clustering of ventricular cardiomyocyte nuclei revealed a spectrum of differentiation from infant, juvenile, and adult stages. We find 488 TFs differentially expressed (p. adj $<$ 0.01, abs. LFC $>$ 1) between two or more timepoints in non-cycling ventricular cardiomyocytes, such as genes continually upregulated across postnatal development such as *Foxo3* and retinoid X receptor gamma (*Rxrg*) (Fig. 3.8, Methods). Several studies have implicated *Foxo3* as a transcriptional regulator of cardiac hypertrophy by inhibiting cardiomyocyte growth and promoting autophagy[210,211], potentially responsible in part for the decreased rate of hypertrophy after PND 14. In the mouse embryo, retinoic acid (RA) signaling establishes polarity and promotes the ventricular phenotype in developing cardiomyocytes[212], therefore *Rxrg* may also be important in maintaining normal ventricular phenotype in the postnatal state. Cardiomyocyte markers such as *Gata4* and *Mef2* family genes, family genes, well-known transcriptional regulators of cardiac genes in infant, juvenile, and adult cardiomyocytes[213–217] are expressed throughout development, highlighting the strong regulatory signature of cardiomyocytes at all ages.

As in the brain, skeletal muscle contains adult stem cells known as satellite cells that continually replenish myonuclei throughout development and adulthood. As muscles grow, quiescent satellite cells characterized by expression of *Pax7* are activated to become proliferating myoblasts[144]. Post-mitotic myoblasts align and fuse with each other to form multinucleated myotubes, expressing myogenic regulatory factors (MRFs) including *Myf5*, *Myod1*, and *Myog*[218,219]. A portion of satellite cells follows an alternative lineage, where they remain unfused and undifferentiated to renew the stem cell pool[218,219]. Myotubes develop further, undergoing structural organization to become mature myofibers with the ability to perform coordinated contraction and relaxation. Mature skeletal muscle fiber types are identified based on the expression of distinct myosin heavy chain proteins. *Myh7* serves as a marker for slow-twitch type 1 fibers, while *Myh2*, *Myh4*, and *Myh1* are specific to fast-twitch type 2 fibers (2A, 2B, and 2X, respectively)[112]. Additionally, *Myh3* has classically been linked

to embryonic fibers, and *Myh8* to perinatal fibers[220]. The gastrocnemius, or calf muscle, extends from two heads attached to the femur and in adults is primarily composed of fast-twitch type 2B fibers which run towards the Achilles tendon[221]. However, fiber type alone provides only a partial understanding of muscle heterogeneity, as the weight of this muscle is sexually dimorphic, with male gastrocnemius weighing 29% more on average than female gastrocnemius at matching timepoints. In our dataset, perinatal myonuclei constitute the majority of myonuclei shortly after birth at PND 4. By PND 10, type 1 myonuclei contribute significantly to the total myonuclei before being surpassed by type 2 fibers, particularly type 2B. However, traces of type 1, as well as type 2A and 2X, persist into adulthood (Fig. 3.1d, 3.9). Among 47 single-nucleus clusters, 6 exhibit a notable difference in proportion between males and females, with 5 myonuclei clusters and 1 fibro-adipogenic progenitor cluster showing a difference exceeding 1 standard deviation from the mean (Fig. 3.9). In addition to tissue-specific cell types, we consistently detect common cell types such as endothelial and immune cells across all our vascularized tissues, maintaining relatively stable proportions. However, their relative proportions in the overall tissue composition varies, with heart tissue having the highest overall counts of endothelial and immune cells (Fig. 3.5, 3.6, 3.7, 3.8, 3.9). In summary, our time course effectively captures dynamics of cell types and cell states during postnatal development.

Topics modeling identifies cellular programs with a core set of regulatory genes

While many genes serve as markers for distinct cell types and states, we hypothesize that cellular programs are fundamentally constructed from a core set of genes including TFs, microRNAs, and chromatin regulators. While a program often controls expression of protein-coding markers that may not be regulators themselves, its core set of regulatory genes governs cell type and state. To study specification of cell types, such as cardiomyocytes, endothelial

cells, and microglia, and transitions between cell states, such as transient adrenal cortex zones, granule cell stages, and muscle fiber types, we applied Latent Dirichlet Allocation (LDA) to our annotated snRNA-seq data in each tissue using Topyfic[128].

LDA is a Bayesian model that learns a limited set of hidden "topics" that can generate the underlying training data[124]. In the context of single-cell RNA-seq, LDA groups genes into topics and assigns them numerical scores or weights based on their relevance to the topic[127,128]. Notably, genes may appear in multiple topics, reflecting the intricate nature of biological systems where genes participate in diverse regulatory programs. Unlike some other methods like WGCNA[122,197], LDA's approach aligns more closely with biological reality[128]. By examining the expression patterns of these weighted genes, LDA assigns a participation score to each cell for each topic, ranging from 0 to 1[128]. A participation score of 1 indicates that a cell's gene expression profile perfectly aligns with the genes associated with that topic[128]. However, it is rare for a cell to participate in just one topic, as numerous cellular processes are affected by regulatory networks[222]. Through the analysis of gene weights, LDA enables the comparison of latent traits associated with topics, offering insights into dynamic cell types and states. Topyfic performs LDA 100 times on a normalized[223] genes-by-cells matrix and determines consensus topics by clustering all 100 runs[100,128]. The resulting set of topics represents the regulatory genes learned to characterize the gene expression profiles in our single cells. These topics can be conceptualized as vectors in gene space, with each weight representing the value in each gene, or dimension. This nuanced approach contrasts with a binary set of marker genes, which merely denotes presence or absence, failing to capture the idea that genes may have multiple roles in different contexts[105,106]. Overall, the topics approach acknowledges the complexity of cellular programs, recognizing that cells likely participate in multiple programs simultaneously, and underscores the diverse roles that genes may play across various functional contexts.

Our approach to identifying cellular programs involves focusing the LDA vocabulary on genes

that we categorize as regulatory. TFs are master regulators of the transcriptome and form the core of cellular programs and gene regulatory networks due to their broad impact on target genes[224]. Despite their significance, TFs exhibit a wide range of expression patterns across different cell types, often being overshadowed by the expression patterns of their target genes[225]. In addition to TFs, genes were selected with GO term annotations that impact transcriptional and chromatin regulation such as chromatin binding genes, transcription regulators, chromatin organizing genes, host genes representing microRNAs, histone modifying genes (acetyltransferases, deacetylases, methyltransferases, and demethylases), and TBP-associated factors as well as members of the Mediator complex (TAF-MED) (Methods). Bulk RNA-seq measurements of these genes by regulatory biotype reveals most variation in TF detection at >1 TPM in at least one bulk sample across tissues (Fig. 3.2a). Out of 1,357 known TFs in the mouse genome, 75% are detected across all tissues, with most in adrenal gland, followed by gastrocnemius and heart, then cortex and hippocampus. Other gene biotypes such as chromatin binding genes, chromatin organizers, and transcription regulators are similarly detected across all tissues (Fig. 3.2b, c, d). Of the smallest categories (microRNA host genes, TAF-MED, and histone modifiers, Fig. 3.2e, f, g), the same pattern of adrenal gland, gastrocnemius, heart, and brain regions appears again in the microRNA host gene category, most likely due to the tissue specificity of microRNA expression[226]. In summary, topics modeling using a curated vocabulary approach aims to extract impactful cellular programs and allows for characterization of regulatory gene biotypes.

Regulatory gene expression is sufficient to define cell types and cell states

To identify topics specific to each cell type within a tissue, we applied Topyfic on each tissue separately, incorporating batch effect correction between snRNA-seq barcoding platforms[128,227]. Selecting the appropriate number of topics, denoted as k, is a crucial aspect of

topic modeling. In our approach, we fine-tuned k within the range of 5 to 35 for each tissue, ultimately settling on the resolution that yielded the same number of topics as the specified value of k. This fine-tuning led to an average of approximately 16 topics per tissue, with the adrenal gland having the highest count at 19, and the hippocampus having the lowest at 14 (Fig. 3.10, 3.11, 3.12, 3.13, 3.14, Methods).

Analysis of topic-trait relationships in hippocampal topics indicates that genes crucial for cell type specification are highly weighted in our topics. Topic-trait relationships are analyzed using Spearman correlations to associate specific topics with traits based on cell participation. We observe that hippocampus topic 1 (HC1) corresponds to astrocytes, HC2 to DG granule cells, HC4 to oligodendrocytes, HC6 to inhibitory GABAergic interneurons, HC10 to OPCs, HC11 to endothelial cells, and HC12 to microglia (Fig. 3.2h). Despite the absence of certain protein-coding genes crucial for cell type-specific functions, such as myelin glycoproteins in oligodendrocytes[196], our identified topics exhibit strong correlations with annotated cell types. Developmental progression through the oligodendrocyte lineage is accompanied by topic switching from HC10 in OPCs, to a mix of HC10 and HC4 in intermediate oligodendrocytes (MFOL) to exclusive enrichment of HC4 in mature oligodendrocytes (MOL). Breakdown of cell participation in OPCs and oligodendrocytes shows gradual expansion of HC4 from 3% to 48% to 85%, while HC10 diminishes from 63% in OPCs to 29% in MFOLs during glial differentiation (Fig. 3.2i). Minor topics HC5 and HC7 remain active throughout differentiation, potentially representing general glial programs that are turned on regardless of subtype. Structure plots are very dense stacked bar plots showing the proportion of topic participation, where each column is a single nucleus grouped by annotated cell type. Ordering of nuclei by pseudotime shows that as cells differentiate, HC10 is gradually replaced by HC4 while minor topics remain constant (Fig. 2i). Notably, topic modeling also captures annotated cell states. HC9 accounts for 43% of the participation of early cells in the DG, while HC2 corresponds to 67% of the participation of mature granule cells (Fig. 2j). Once again, ordering by pseudotime emphasizes topic switching, as HC9 decreases during granule

cell maturation. Thus, the expression patterns of regulatory genes alone suffices to define both transcriptional cell types and cell states.

Comparing the number of topics detected per nucleus, we observed that most nuclei in each tissue are effectively characterized by more than one topic, and a median of 2 topics accounts for 80% of cell participation (Fig. 3.3a). This result supports our hypothesis that cells concurrently run multiple programs, especially during transitional processes of differentiation or maturation[194], as evidenced here in hippocampal cell types. Importantly, topics with high cell participation are consistently enriched for specific cell types and states, a trend observed across all tissues (Fig. 3.3b, 3.10, 3.11, 3.12, 3.13, 3.14). Conversely, topics with low participation are typically not associated with any particular cell type (Fig. 3.3b, shaded gray). At our chosen resolution, all cell types with >1,400 nuclei are captured by at least one topic. In addition to having the highest number of topics compared to other tissues, adrenal gland has the most distinct annotated cell types (10), surpassing other tissues (6, 7, 8, and 8 in cortex, hippocampus, heart, and gastrocnemius, respectively). Interestingly, in both the adrenal gland and heart, a particular topic consistently showed enrichment in cycling cells, irrespective of their cell type of origin (Fig. 3.10, 3.13).

Tissue-specific signals in microglia and macrophage topics

Immune cells are represented by topics with high cell participation across all five tissues. In cortex and hippocampus, topics CX8 and HC12 are associated with microglia, while AD14, HT3, and GC10 correspond to resident macrophages in the adrenal gland, heart, and gastrocnemius, respectively (Fig. 3.3a). Microglia, the brain's resident immune cells, originate from progenitors formed during the first wave of primitive hematopoiesis around embryonic day (E) 7.5[228,229]. They migrate to the developing central nervous system (CNS) through the bloodstream, typically around E9.5 in mice[230]. After prenatal establishment in

the CNS, microglia undergo proliferation and expansion, reaching their peak two weeks after birth and sustained through low proliferation levels into adulthood[230]. The second wave of hematopoiesis gives rise to yolk sac macrophages, a portion of which expand and differentiate into tissue-resident macrophages by E9.5[228]. While previous studies have compared the gene expression profiles of macrophages and microglia derived from adult human brain and blood in culture[231,232], as well as infiltrating macrophages and microglia in adult rat brain[233], our approach leverages multiple coordinated tissues from the same individual mice.

An MA plot of gene weights for the microglial topics in hippocampus (HC12) vs cortex (CX8) reveals very similar topic compositions, aligning with our expectations (Fig. 3.3c). Very few genes have an absolute log ratio (M) value > 5 (47 in hippocampus, 8 in cortex) (Fig. 3.3c), none of which have been implicated in regional microglial signatures. Genes involved in microglia polarization (e.g., *Irf8*[234] and *Stat3*[235]), activation and inflammatory response (e.g., *Spi1*[236] and *Irf2*[237]), and establishment of microglia identity and immune response (e.g. *Sall1*,[238] *Sall3*[239], *Etv5*[240], and *Zeb1*[241] all have high mean average (A) values in both cortex and hippocampus microglia topics. Thus, regulatory topics assign similar weights for genes from identical cell types in different tissues when trained independently.

By contrast, comparison of hippocampus microglia topic HC12 and heart macrophage topic HT3 reveals 165 genes with $|M| > 5$ (67 in hippocampus, 98 in heart). Microglia-specific genes such as *Sall1*, *Sall3*, *Etv5*, and *Zeb1* are more highly weighted in hippocampus, whereas genes involved in macrophage differentiation, polarization, and inflammatory pathway signaling such as *Runx3*[242], *Foxo1*[243,244], and *Tfec*[245,246] exhibit higher weights in heart (Fig.3.3d). Interestingly, *Tfec* expression has been shown to be activated by Stat6, another heart-specific macrophage TF in our comparison, which transduces IL-4 signals and binds to the promoter of *Tfec*[246] (Fig. 3.3d). Additionally, *Foxo1* expression has been linked to cardiac fibrosis following macrophage activation[247]. Due to their similar weights across topics in both tissues, *Spi1*, *Irf2*, *Irf8*, and *Stat3* may belong to a common transcriptional signature of shared

immune functions between postnatal microglia and macrophages.

Mitosis topics are driven by chromatin regulators

We then asked whether particular classes of regulatory genes were found in most topics or were more specific to a subset of topics. We calculated the percentage of topics where a gene surpasses a minimal weight threshold of 1 compared to the median of its weight across all topics (Fig. 3.3e-k). Notably, 30% or more of genes classified as chromatin regulators (Fig. 3.3f, h, j) occupy the upper right quadrant, indicating they are highly weighted in most topics. In contrast, transcription factors, transcription regulators, microRNA host genes, and the TAF and Mediator complex family of genes exhibit a different pattern, with 20% or less highly weighted in most topics (Fig. 3.3e, g, i, k). TFs are mostly either highly weighted and topic-specific (59%, upper left quadrant) or specific with lower weights (26%, lower left quadrant). A simplified analysis of gene biotype enrichment within topics revealed two topics (HT6 and AD5) highly enriched for chromatin regulators compared to TFs and microRNA host genes (Fig. 3.3j, Methods). Interestingly, these topics correspond to our cycling topics, primarily influenced by a proliferative state rather than their cell type of origin (Fig. 3.10, 3.13). Our results suggest that cellular programs essential for mitosis, particularly those governing chromatin condensation and structure, are primarily orchestrated by chromatin regulators. In contrast, programs driven by transcription factors play a lesser role in directing a proliferative cell state.

Topics in shared cell types from diverse tissues cluster together

Cosine similarity between topics serves as a measure to compare gene weights, representing the angle between two topics in gene space. It is similar to other correlation methods, where 0 indicates low concordance between topics and 1 represents high concordance. By comput-

ing the cosine similarity for each pair of topics among the 82 total topics, and subsequently filtering clusters for those with a cosine similarity above 0.9, we identified 20 distinct clusters of topics (Fig. 3.3m, Methods). As expected, cycling topics HT6 and AD5 are highly correlated with a cosine similarity of 0.93, along with a large cluster of endothelial topics across all five tissues (C9 and C1, respectively, Fig. 3.3m). Topics representing common cell types across brain regions cluster in C4 (glutamatergic neurons), C10 (GABAergic interneurons), C11 (microglia), C12 (astrocytes), C13 (OPC), and C14 (oligodendrocytes). Interestingly, the macrophage cluster C3 is distinct from the microglia cluster C11. As observed in comparing HC12 and HT3 (cosine similarity 0.83, Fig. 3.3d), tissue-specific signatures in macrophages and microglia likely drive the differences in gene weights between microglia and macrophage topics. C1 includes two cardiac heart topics, while C19 and C20 represent additional signatures in cardiac endothelial and endocardial cells, distinct from the general endothelial signature shared across all five tissues. In summary, the regulatory topics capture core cellular programs that can be compared across tissues with related cell types.

Characterization of cell type specificity in candidate cis-regulatory elements

TFs regulate expression of target genes by binding to cis-regulatory elements (CREs) in open chromatin[225]. The landscape of open chromatin, measured using single nucleus ATAC-seq, provides insight into accessible regulatory elements at the single-cell level. We leveraged the ENCODE registry of candidate cis-regulatory elements (cCREs) in mouse derived from chromatin accessibility, histone modifications, and DNA affinity purification sequencing[248] to score our snATAC-seq data across a cohesive set of chromatin regions. These elements play crucial roles in gene regulation by providing binding sites for transcription factors and influencing chromatin accessibility[248]. Around 43% of these regions are classified as candidate

distal enhancers by H3K27ac and DNase I hypersensitivity, 12% as proximal enhancers, and 31% were determined by chromatin accessibility data alone (Fig. 3.15). Accessibility across the full set of 926,843 cCREs was scored in pseudobulk snATAC nuclei using the integrated clusters from snRNA-seq analysis. The cCREs >5 RPM in at least one pseudobulk cluster per annotated cell type (390,146 total across our tissues) were classified as specific, shared, general, or global by mapping each cluster to its annotated cell type. We categorized cCREs accessible in only one cell type as 'specific', those accessible in more than one cell type within or across tissues as 'shared', those accessible in all major cell types within a tissue as 'general', and cCREs accessible in all major cell type across all tissues as 'global'. Most cCREs are either specific to one cell type (43.1%) or shared (47.9%), with only 9% classified as general or global (Fig. 3.4a). The cell-type-specific landscape of accessible regulatory elements, particularly enhancers, sets the stage for transcription factors to bind and dynamically control gene expression during postnatal development.

Tissue-specific analysis reveals the most cell type-specific elements in cerebral cortex and hippocampus, driven by robust neuronal signatures, with the heart displaying the least cell type specificity (Fig. 3.4b). Indeed, breakdown by cell type in the hippocampus emphasizes glutamatergic neurons as the most specific, and to a lesser extent microglia and pericytes (Fig. 3.4c). In other tissues, the major cell type also exhibits a robust chromatin signature, such as myonuclei in the gastrocnemius and cortical cells in the adrenal gland (Fig. 3.4d, e). To further explore the dynamics and sex specificity of the chromatin landscape, which likely contribute to variations between certain cell types, differential expression analyses were conducted between timepoints and sexes in accessible cCREs. The largest proportion of differentially accessible cCREs between PND 14 and 2 months, are detected in gastrocnemius tissue, while most sex-differential cCREs are detected in adrenal gland (Fig. 3.4f). These results likely reflect the ongoing biological processes within the major cell types of these tissues; myonuclei in the gastrocnemius are transitioning to their mature fiber type, and the emergence of the X-zone in the adrenal zona fasciculata occurs during puberty, emphasizing

the dynamic nature of chromatin accessibility during crucial postnatal stages.

Regulatory motifs are enriched in cell-type-specific cCREs

Although most perinatal myonuclei disappear by PND14, type 1 fibers and fibro-adipogenic progenitors recede while 2B fibers expand, ultimately constituting over three-quarters of the nuclei in gastrocnemius by 2 months (Fig. 3.9). Given that the majority of dynamic cCREs are cell-type specific (Fig. 3.4g), and the predominant cell-type-specific cCREs are found in myonuclei (Fig. 3.4d), we focused on TF binding in myonuclear subtypes. We performed motif enrichment analysis using ArchR[249] in myonuclei-specific cCREs broken up by accessibility in muscle fibers and satellite cells to identify potential regulators which can then be matched to TFs featured in our topic modeling (Methods, Fig. 3.16). Notably, some TFs exhibit concordant motif activity patterns and topic weight. The *Pax7* motif is both enriched in satellite-specific cCREs (Fig. 3.4h) and included in the satellite-associated topics (Fig. 3.4i). Alternatively, *Myog* binding is detected and the TF is highly weighted in one major satellite topic (GC15, 44% participation in satellites), whereas it is not detected in the minor satellite topic (GC8, 12% participation) (Fig. 3.4h,i, 3.14). The more dominant topic potentially reflects satellite cells undergoing postnatal myogenesis, while the minor topic may signify the self-renewing pool of satellite cells actively inhibiting the expression of MRFs[218,219]. Previous studies have found interactions between *Tcf12* and *Mef2c* and MRFs such as *Myod1* in skeletal muscle implicated in skeletal muscle formation[250–254]. While *Tcf12* is weighted in nearly all myonuclear topics, motif enrichment shows activity restricted to satellite cells, in which it has shown to be a crucial regulator of their chromatin remodeling[250]. Similarly, *Mef2c* is found in all non-satellite topics but active only in type 1 myonuclei. *Mef2c* has been linked to type 1 specification by responding to calcium-dependent signaling pathways to promote the transition between fast glycolytic fibers to slow oxidative fibers[251–254]. These observations may reflect cases where target genes are inaccessible while

the TF continues to be expressed.

Comparison of sex-specific regulatory activity in the adrenal zona fasciculata

We then turned to sex-specific cCREs are also celltype-specific in adrenal gland (Fig. 3.4j). Unsurprisingly, female cCREs overlap those attributed to the X-zone and zona fasciculata (Fig. 3.4k), as well as adipocytes. In males, a faint signature is seen in the nuclei annotated as male ZF. We focused motif enrichment on the X-zone, male ZF, and non-sex-specific ZF to investigate binding activity of key TFs from differential expression analysis and topics modeling. *Runx2*, upregulated in female compared to male ZF, has distinct binding activity in X-zone-specific cCREs (Fig. 3.4l). It is also a top-weighted gene in the X-zone topic AD6 (Fig. 3.4m). Despite a previous study in *Runx2* knockout mice suggesting no direct contribution to sex determination[255], it may regulate genes involved in steroid metabolism, as evidenced in mouse osteoprogenitor cells[256]. Furthermore, estrogen receptor alpha has been observed to colocalize with Runx2 in breast cancer and osteoblasts, although their expression is inversely related[257]. In contrast to *Runx2*, although *Thrb* is also differentially upregulated in female ZF but is weighted similarly in X-zone topic AD6 and male ZF topic AD12 with binding activity solely in the male ZF (Fig. 3.4l, m). Likewise, the androgen receptor Ar is highly weighted in both the X-zone topic AD6 as well as male ZF topic AD12, but only active in male ZF (Fig. 3.4l, m). *Ar* is expressed in both male and female sex-specific regions, although more so in the X-zone compared to the male-specific ZF (Fig. 3.5). Recent studies have identified androgen signaling via the androgen receptor as a requirement for X-zone regression during puberty in male mice[258], while *Ar* signaling is not essential for regression in female mice[259]. Our results suggest androgen signaling in male ZF may be mediated by lower levels of *Ar* compared to female ZF, perhaps due to co-activator expression, accessible chromatin at target gene promoters, or involvement of factors from other tissues, such as the

hypothalamic-pituitary-gonadal axis. More broadly, the sexual dimorphic binding activity of transcription factors that are similarly expressed in these homologous cells highlights the fundamental limitations of studying gene regulation using RNA expression alone when ignoring sex as a biological variable.

3.4 Discussion

The ENCODE4 mouse single-nucleus dataset stands out from other genomic catalogs by offering a comprehensive map of postnatal development across diverse tissues, spanning from just after birth to late adulthood. Its strength is further evident in the inclusion of both sexes at each timepoint, setting it apart from datasets like *Tabula Muris Senis*, which is limited to one sex at certain timepoints. This inclusivity allows us to analyze sexual dimorphism across time, such as the emergence of sex-specific adrenal cortex populations during puberty. The dataset facilitates comparisons of maturation rates across tissues, revealing significant differences. For instance, the most significant changes in the adrenal gland occur between 2 months and 18-20 months as sex-specific cortical layers regress, while the largest changes in gastrocnemius occur from postnatal day 4 to postnatal day 10 as myofibers mature. This high-resolution timecourse enables investigations into large-scale dynamics as well as the maintenance of adult stem cell pools like OPCs, NPCs, and satellite cells. Additionally, the integration of snRNA-seq data between Parse and 10x barcoding platforms underscores the complementary information captured by each technology. In summary, this dataset presents a unique opportunity to explore postnatal development throughout the entire mouse body at unprecedented single-cell resolution, offering insights from various biological and technical perspectives.

All experiments were conducted in a B6/CAST hybrid genotype, facilitating future exploration of the genetic basis of complex molecular traits. B6J (*M. m. domesticus*), which

is the most commonly used laboratory mouse and the first to have its genome published, diverged from CAST (*M. m. castaneus*) approximately one million years ago[30,56]. As a wild-derived strain, CAST harbors 17.6 million single-nucleotide polymorphisms compared to the B6J reference genome[48] and exhibits differences in phenotypes such as temperament and hearing ability[260]. These strains represent broader genetic diversity, resembling natural populations, and are two of the founders of the Collaborative Cross[64]. An open question is whether any of the cell states described here would be specific to the F1. Examining gene expression differences in both B6 and CAST parents with our results in the offspring would allow us to determine the impact of a particular allele as acting in *cis* or *trans*. Besides allele-specific gene expression, we could also compare traits such as proportions and dynamics of cell types, as well as participation in the regulatory topics described here, streamlining the identification and analysis of cell types and states.

We applied Topyfic to integrated combinatorial barcoding and multiome datasets, focusing on a curated vocabulary of 2,701 regulatory genes. This analysis revealed 82 regulatory topics associated with 46 distinct cell types and states. Our results indicated an enrichment of transcription factor (TF) and microRNA gene biotypes in cell-type-specific topics, while cycling topics are predominantly influenced by chromatin regulators. Although most studies of polyadenylated RNA ignore the impact of microRNAs, a significant fraction of microRNAs are intragenic, most of which are found within introns of protein-coding genes[261,262]. MicroRNAs can be transcribed by RNA polymerase II together with their host genes[263], suggesting that cell-type markers may have microRNAs embedded in their introns, potentially playing a major role in the transcriptional regulation of that cell type. Additionally, our analysis identified correlated regulatory topics across tissues for common cell types, such as endothelial cells, while immune cells retained a tissue-specific signature, particularly in trunk organs compared to brain microglia. We further classified ENCODE v4 cCREs based on accessibility in our cell types, revealing that nearly half of the identified cCREs exhibit cell-type specificity. Lastly, we explored motif enrichment patterns of TFs within topics in

cell type- and state-specific regulatory elements.

The behavior of rLDA topics aligns with our hypotheses regarding genuine cellular programs: they are predominantly cell type- and state-specific, often co-expressed, reproducible across tissues, and can be defined using regulatory genes alone, especially TFs. Focusing on regulatory genes offers a more direct insight into cellular programs by ensuring the inclusion of TFs in each topic. It's crucial to note that a TF's presence in a topic doesn't automatically imply active involvement in regulatory programs, and further verification may require follow-up experiments and integration with chromatin accessibility or DNA binding data. By leveraging corresponding chromatin accessibility data, we identified cases where a top-weighted TF exhibits enriched binding in a cell type associated with its topic, as well as instances where topic TFs are active in different cell types or states. Our results demonstrate the successful identification and interpretation of cellular programs using topic modeling across multiple tissues and barcoding platforms, establishing a foundational understanding of transcriptional programs in the developing mouse.

Acknowledgements

We thank the Caltech Jacobs Genetics and Genomics Laboratory for sequencing the bulk mRNA-seq libraries. Ali Mortazavi and Barbara J. Wold were supported by UM1HG009443. B.J.W. was also supported by the Caltech Beckman Institute BIFGRC. Jennifer Jou and Ingrid Youngworth (ENCODE DCC) were supported by U24HG009397.

Author contributions

All the work presented here from experimental design to writing the manuscript was guided extensively by Barbara J. Wold and Ali Mortazavi. Narges Rezaie performed data processing, data analysis, generated figures, wrote and edited the manuscript. Brian Williams performed bulk mRNA-seq experiments and wrote and edited the manuscript. Annika Weimer, Minyi Shi, and Xinqiong Yang performed 10x Multiome experiments. Heidi Liang performed Parse

Biosciences snRNA-seq experiments and sequenced snRNA-seq and microRNA-seq libraries. Louise Dionne and Laura Reinholdt bred mice, dissected tissues, and shipped samples to Caltech. Samuel Morabito performed data processing and Fairlie Reese edited the manuscript. Fairlie Reese, Diane Trout, Jennifer Jou, and Ingrid Youngworth contributed to ENCODE uniform processing pipeline development for Parse Biosciences snRNA-seq data. All authors read and approved the final manuscript.

Data and code availability

- Data availability: ENCODE carts of all data used are listed in Table S1.
- Data processing/figure generation code

3.5 Supplementary tables

- Table S1: ENCODE data carts listing all data used.
- Table S2: snRNA-seq and snATAC-seq metadata and QC filtering thresholds.
- Table S3: List of marker genes used in cell type annotation.
- Table S4: Gene weights in adrenal gland topics.
- Table S5: Gene weights in left cerebral cortex topics.
- Table S6: Gene weights in hippocampus topics.
- Table S7: Gene weights in heart topics.
- Table S8: Gene weights in gastrocnemius topics.

3.6 Methods

Mice and tissue collection

All animals were treated and housed in accordance with the Guide for Care and Use of Laboratory Animals. Approval for all experimental procedures was granted by Caltech's Institutional Animal Care and Use Committee (IACUC), aligning with both institutional and national guidelines. Samples were obtained from animals covered under the approved IACUC protocol #IA21-1647, "Single-cell transcriptome studies from multiple mouse tissues". Tissues at postnatal day (PND) 4, PND 10, PND 14, PND 25, PND 36, 2 months, and 18-20 months from C57BL6/J (RRID:IMSR_JAX:000664) × CAST/EiJ (RRID:IMSR_JAX:000928) F1 hybrid mice were obtained from Jackson Laboratories (JAX). Adrenal gland and gastrocnemius tissues were pooled from 3 individuals for PND 4 and PND 10 timepoints. Hippocampus tissues were pooled from 3 individuals for PND 10 and PND 14 timepoints. Tissues were flash-frozen in liquid nitrogen and delivered to Caltech on dry ice, where they were stored at -80°C until RNA extraction.

Isolation of RNA for bulk assays

For bulk RNA-seq, total RNA was extracted from flash-frozen tissues at Caltech using the Norgen Animal Tissue RNA Purification Kit (Norgen Biotek cat. #25700). The tissue was lysed using Buffer RL and proteins were digested with proteinase K. Genomic DNA was removed with DNaseI treatment on the columns. The purified total RNA includes includes large mRNAs, lncRNAs, and small RNAs. The Qubit dsDNA HS Assay Kit (Thermo cat. #Q32854) was used to assess RNA concentration and RIN values were determined using the Bioanalyzer Pico RNA kit (Agilent cat. #5067-1513), with average RIN scores of 8.2 for the adrenal gland, 9.1 for the hippocampus, 9.3 for the cortex, 9.0 for the heart, and 9.3 for

gastrocnemius tissues.

Bulk RNA-seq from mouse tissues

Each cDNA library was built from 300 ng total RNA with ERCC spike-ins (Thermo cat. #4456740) using the NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (NEB cat. #E7760), specifically the protocol for use with NEBNext Poly(A) mRNA Magnetic Isolation Module (NEB cat. #E7490). Ribosomal RNA was depleted from total input RNA using the NEBNext rRNA Depletion Kit (NEB cat. #E6310). First and second strand synthesis, cDNA end prep, adapter ligation, and finally PCR amplification resulted in the final libraries. The libraries were quantified using the Qubit dsDNA HS Assay Kit (Thermo cat. #Q32854) and sequenced on an Illumina HiSeq 2500 as 100 bp single-end reads to 50 M raw read depth. For submission to the ENCODE portal, libraries needed at least 30 M aligned reads and a Spearman replicate correlation >0.9.

Purification of nuclei for Split-seq

For Parse Split-seq experiments performed at UCI, nuclei were isolated from the 5 core tissues (adrenal gland, left cerebral cortex, hippocampus, heart, and gastrocnemius) for all 7 timepoints (PND 4, PND 10, PND 14, PND 25, PND 36, 2 months, and 18-20 months). Flash-frozen tissues shipped from Caltech were transferred to a chilled gentleMACS C Tube (Miltenyi Biotec cat. #130-093-237) with 2 mL Nuclei Extraction Buffer (Miltenyi Biotec cat. #130-128-024) supplemented with 0.2 U/uL RNase Inhibitor (NEB cat. #M0314L) on ice. Nuclei were dissociated from whole tissues using a gentleMACS Octo Dissociator (Miltenyi Biotec cat. #130-095-937). Suspensions were filtered through a 70 um strainer then a 30 um strainer (Miltenyi Biotec cat. #130-110-916 and #130-098-458, respectively). Nuclei were resuspended in cold PBS + 7.5% BSA (Life Technologies cat. #15260037) and 0.2

U/ul RNase inhibitor for manual counting using a hemocytometer and DAPI stain (Thermo cat. #R37606). For gastrocnemius tissue, debris was removed from nuclei suspensions with Debris Removal Solution (Miltenyi Biotec cat. #130-109-398). Nuclei were mixed with Debris Removal Solution and layered on top of PBS, then centrifuged at 4°C, 3000 x g for 10 minutes with full acceleration and no brake. Nuclei bands were separated from debris layers and concentrations were determined using a hemocytometer. For Parse Split-seq, 1-4 million nuclei per sample were fixed using Parse Biosciences' Nuclei Fixation Kit v1 (Parse Biosciences cat. #WN100), following the manufacturer's protocol. Briefly, nuclei were incubated in fixation solution for 10 minutes on ice, followed by permeabilization for 3 minutes on ice. The reaction was quenched, then nuclei were centrifuged and resuspended in 300 uL Nuclei Buffer (Parse Biosciences cat. #WN101) for a final count. DMSO (Parse Biosciences cat. #WN105) was added before freezing fixed nuclei at -80°C.

Parse Split-seq experiments

Nuclei were barcoded using Parse Biosciences' Evercode WT Kit v1 (cat. #EC-W01030), following the manufacturer's protocol. Briefly, fixed nuclei were thawed and added to the Round 1 reverse transcription barcoding plate at 15,000 nuclei per well across 48 wells. Individual samples from each tissue were distributed in sample barcoding plates with at least 1 well per individual. Within the fixed nuclei, RNA was reverse transcribed using oligodT and random hexamer primers and the first barcode was annealed. After RT, nuclei were pooled and distributed in 96 wells of the Round 2 ligation barcoding plate for in situ barcode ligation. After Round 2, nuclei were pooled and redistributed into 96 wells of the Round 3 ligation barcoding plate for barcode 3 and Illumina adapter ligation. Finally, nuclei were counted using a hemocytometer and distributed into 6 subpools for adrenal, 6 subpools for cortex, 5 subpools for hippocampus, 4 subpools for heart, and 5 subpools for gastrocnemius, each containing 12,000 nuclei, with 2 additional subpools of 15,000 nuclei for gastrocnemius.

Nuclei from each tissue were also distributed into 1-2 small subpools of 1,000-2,000 nuclei each, for a target of around 75,000 nuclei per tissue (>500 UMI). The nuclei in each subpool were lysed and the barcoded cDNA underwent template switching and amplification. The cDNA was cleaned using AMPure XP beads (Beckman Coulter cat. #A63881) and quality checked using the Qubit dsDNA HS Assay Kit (Thermo cat. #Q32854) and a Bioanalyzer 2100 (Agilent cat. # G2939A) High Sensitivity DNA Kit (Agilent cat. #5067-4626) before proceeding to Illumina library preparation with 100 ng of full-length cDNA per subpool. Subpool cDNA was fragmented and Illumina P5/P7 adapters were ligated during the final amplification, followed by size selection and quality check with the Bioanalyzer and Qubit. Libraries with 5% PhiX spike-in were sequenced on an Illumina NextSeq 2000 sequencer with P3 200 cycles kits (Illumina cat. #20040560) as paired-end, single-index reads (115/86/6/0) to an average depth of 181 M reads per 12,000-15,000-nucleus library and an average depth of 134 M reads per 1,000-2,000-nucleus library.

Purification of nuclei for 10x Multiome

For 10x Multiome experiments performed at Stanford University, nuclei were isolated from 5 core tissues for PND 14 and 2 month timepoints. Flash-frozen tissues were dissociated in a Douce homogenizer with 1 mL homogenization buffer: 0.26 M sucrose (Sigma cat. #S7903-250G), 0.03 M KCl (Thermo cat. #AM9640G), 0.01 M MgCl2 (Thermo cat. #AM9530G), and 0.02 M Tricine-KOH pH 7.8 (Sigma cat. #T0377), supplemented with 0.6 U/uL RNase Inhibitor (Thermo cat. #EO0384). Suspensions were filtered through a 40 um strainer (Fisher Scientific cat. #22363547) and debris was removed using an iodixanol gradient. Iodixanol solution was diluted from 60% iodixanol (Sigma cat. #D1556-250ML) with dilution buffer consisting of 0.15 M KCl, 0.03 M MgCl2, and 0.12 Tricine-KOH pH 7.8. Nuclei were mixed 1:1 with 50% iodixanol solution, then 30% iodixanol solution was layered underneath the 25% mixture, and 40% iodixanol solution was layered at the bottom. Nuclei were

centrifuged at 4°C, 3000 x g for 20 minutes with full acceleration and no brake and the nuclei band was separated from the debris layer. Concentrations of the final suspensions were determined using a hemocytometer. Nuclei were immediately processed following the Chromium Next GEM Single Cell Multiome ATAC + Gene Expression User Guide (CG000338).

10x Multiome experiments

Gene expression and chromatin accessibility were profiled simultaneously in the same nuclei using the Chromium Next GEM Single Cell Multiome ATAC + Gene Expression kit (10x Genomics cat. #1000283) following the manufacturer's protocol. Briefly, around 16,000 nuclei were loaded per well in the microfluidic chip and partitioned into gel beads-in-emulsions (GEMs) for a target recovery of 5,000-10,000 nuclei per sample (around 80,000 nuclei per tissue). During incubation, transposase cleaved open regions of DNA and added GEM-specific adapter sequences to the fragments. After transposition, the nuclei lysates were reverse transcribed using oligodT primers, which also adds GEM-specific barcodes and UMIs to the resulting cDNA. The GEMs were then broken and the transposed DNA and barcoded cDNA underwent pre-amplification PCR to produce the input material for parallel snATAC-seq and snRNA-seq library building. For snATAC-seq, Illumina P5/P7 adapters were added during sample index PCR and the final libraries were cleaned using SPRIselect beads (Beckman Coulter cat. #B23318). For snRNA-seq, the barcoded cDNA underwent template switching and amplification, and was then fragmented and size-selected using SPRIselect beads. Illumina P5/P7 adapters were added during sample index PCR and the final snRNA-seq libraries were cleaned using SPRIselect beads. The snATAC-seq libraries were sequenced on an Illumina NovaSeq 6000 sequencer as paired-end, dual-indexed reads (50/50/8/24) to an average depth of 180 M reads per library. The snRNA-seq libraries were sequenced on an Illumina NovaSeq 6000 sequencer as paired-end, dual-indexed reads (28/90/10/10) to an average depth of 194 M reads per library.

Demultiplexing Parse Biosciences snRNA-seq data

Due to the combinatorial barcoding approach, raw fastqs from Parse snRNA-seq libraries contain all samples included in the experiment. In order to provide sample-level fastqs to the ENCODE portal, Parse Biosciences' split-pipe software v0.7.6p and custom code were used to assign reads to samples. Briefly, split-pipe v0.7.6p was used to generate an annotated fastq with read names containing cell barcodes (process/single_cells_barcoded_head.fastq.gz) as well as a cell metadata file (all-well/DGE_unfiltered/cell_metadata.csv) mapping barcode to sample for each pair of subpool fastqs associated with an experiment. A custom python script calls seqtk v. 1.3-r106 (`https://github.com/lh3/seqtk`) to extract reads from the original fastqs and output them as sample-level fastq files.

Read mapping and quantification

All data quantifications were downloaded from ENCODE portal using carts, organizing the data based on assay and/or tissue (refer to Table S1 for links to carts).

Bulk and single-nucleus RNA-seq data were processed through ENCODE uniform processing pipelines using the mm10 genome with Gencode vM21 annotations. For bulk RNA-seq, the data were aligned using STAR v. 2.5.1b[176] and quantified using RSEM, which provides FPKM, TPM, and raw counts (`https://www.encodeproject.org/pipelines/ENCPL862USL/`).

The snRNA-seq data were aligned using STARSolo v. 2.7.10a[94] with GeneFull_Ex50pAS settings to generate UMI count matrices (`https://www.encodeproject.org/pipelines/ENCPL257SYI/`), similar to the intronic count option in 10x's Cell Ranger. Single-nucleus ATAC-seq data were processed using the standard ENCODE snATAC-seq pipeline with the mm10 genome to generate fragment files which were used as input to downstream analyses

(https://www.encodeproject.org/pipelines/ENCPL952JRQ/).

Bulk RNA-seq analysis

Normalized bulk RNA-seq quantifications were concatenated across all samples using the TPM column from the ENCODE pipeline. In each tissue, the number of regulatory genes in each category were counted if they were expressed at >1 TPM in at least 1 bulk sample.

QC and filtering of single-nucleus data

Analyses were performed on a per-tissue basis and all input files were downloaded from the ENCODE portal. The snRNA-seq tar files contain sparse matrices with corresponding gene and barcode CSV files. The corresponding snATAC-seq tar files for 10x Multiome contain compressed TSV fragments and indices. For Parse Split-seq, the number of datasets varies depending on the number of subpools set aside per tissue.

To perform the integrated snRNA-seq analysis, 42 Parse Split-seq datasets and 8 10x Multiome datasets for adrenal gland, 32 Parse Split-seq datasets and 8 10x Multiome datasets for cortex, 34 Parse Split-seq datasets and 8 10x Multiome datasets for hippocampus, 28 Parse Split-seq datasets and 8 10x Multiome datasets for heart, and 56 Parse Split-seq datasets and 8 10x Multiome datasets for gastrocnemius were downloaded from the ENCODE portal (Table S3). Genes were filtered for protein coding, lncRNAs, pseudogenes, and microRNAs. Ambient RNA was filtered from droplet-based 10x data using Cellbender v. 0.2.2[264]. Doublet detection was performed on nuclei with > 500 UMIs detected per nucleus using Scrublet v. 0.2.3[97].

Data were filtered differently for the "standard" Parse Split-seq libraries (12,000-15,000-nucleus subpools), small Parse Split-seq libraries (1,000-2,000-nucleus subpools), and 10x

Multiome nuclei (5,000-nucleus libraries). The Parse Split-seq nuclei belonging to the 12-15,000-nucleus subpools were filtered by $>$ 500 and $<$ 30,000 UMIs per nucleus, $>$ 500 genes expressed, $<$ 0.2 doublet score, and $<$ 0.5 percent mitochondrial gene expression for adrenal gland, cortex, and hippocampus, and the 1-2,000-nucleus subpools by $>$ 1000 and $<$ 50,000 UMIs. For heart, the filters were relaxed slightly to $<$ 0.25 doublet score and $<$ 1 percent mitochondrial gene expression and further relaxed for gastrocnemius to $<$ 5 percent mitochondrial gene expression. The 10x Multiome nuclei were filtered slightly differently: $>$ 500 and $<$ 30,000 UMIs, $>$ 300 genes, $<$ 0.25 doublet score, and $<$ 5 percent mitochondrial gene expression for cortex, hippocampus, and gastrocnemius, and $>$ 1000 UMIs, $<$ 0.2 doublet score, and $<$ 0.5 percent mitochondrial gene expression for adrenal gland and heart. In addition, 10x Multiome nuclei were also filtered by $>$ 1000 unique nuclear fragments, TSS enrichment $>$ 4, and $<$ 1 ArchR doublet score in the corresponding snATAC-seq data. After initial processing of snATAC-seq data (described below), barcode sequences from snRNA-seq and snATAC-seq multiome nuclei were matched and nuclei failing snATAC-seq QC were excluded from downstream snRNA-seq analysis. All filtering parameters per library can be found in Table S2.

Preprocessing 10x snATAC-seq data

ArchR Arrow files were generated for each tissue using the ENCODE processed fragments files from 8 experiments with a minimum TSS enrichment of 4, minimum 1,000 unique fragments per cell, and excluding reads from mitochondrial DNA in downstream analysis[249]. Doublets were scored and filtered using ArchR's "addDoubletScores" and "filterDoublets" functions with an enrichment threshold of 1[249]. ArchR projects for each tissue were saved and barcode sequences were translated into their snRNA-seq counterpart and saved as csv files. After snRNA-seq filtering, nuclei failing snRNA QC were dropped from the ArchR project using "subsetArchRProject".

Integration of Parse and 10x snRNA-seq data

After filtering the 3 Seurat objects per tissue (standard Parse, small Parse, and 10x Multiome), each was normalized using the function "SCTransform" in Seurat v. 4.1.1[98], with number of genes expressed per nucleus and percent mitochondrial gene expression regressed out. Anchors for integration across the 3 objects were calculated using "SelectIntegrationFeatures" with 3,000 genes, "PrepSCTIntegration", and "FindIntegrationAnchors" in Seurat, with the standard Parse dataset serving as the reference due to inclusion of all 7 timepoints. After integrating data ("IntegrateData"), principal component analysis was performed on the integrated assay by the "RunPCA" function with 50 principal components, with the UMAP ("RunUMAP") calculated from the first 30 components. Clustering was performed with the Louvain clustering algorithm ("FindClusters") with resolution 0.8, with sub-clustering performed as necessary on specific clusters in gastrocnemius and hippocampus due to expression of known marker genes (Fig. 3.7, 3.9).

Integrated cell type annotation

When available, reference datasets were used to transfer annotations using "FindTransferAnchors" in Seurat v. 4.1.1[98]. For both cortex and hippocampus, a downsampled version of the 1M whole cortex and hippocampus 10x atlas from 8 week old mice available on the Allen data portal[196] was used to transfer subtype-level annotations. Downsampling was performed per "cell_type_alias_label" group, with 1,000 nuclei taken per cell type (or all nuclei, if $< 1{,}000$ were available) for a total of 250,734 nuclei used for label transfer. For the heart dataset, both a human heart cell atlas[198] (486,134 nuclei) and a dataset of 8-14 week old stressed mouse ventricles[199] (29,615 nuclei) were used separately for label transfer. For gastrocnemius, label transfer was performed using P10, P21, and 5-month mouse tibialis anterior datasets[112] (28,047 total nuclei). In addition to label transfer, curated marker genes

were used to refine predictions (Fig. 3.5, 3.6, 3.7, 3.8, 3.9, Table S3). In lieu of a reference dataset in the case of adrenal gland, marker genes alone were used to annotate celltypes per cluster (Fig. 3.5, 3.6, 3.7, 3.8, 3.9, Table S3). Each cluster was annotated at the finest possible resolution in a grouping titled "subtypes" (in all figures, metadata, and data objects). This resolution includes dynamic cell states such as OPCs, early DG, the sex-specific populations in the adrenal cortex, and layer-specific neuronal subtypes in cerebral cortex. Depending on the downstream analysis, subtypes and states were grouped into a coarser resolution titled "celltypes". For example, transient sex-specific populations in the adrenal cortex are collapsed along with zona fasciculata, and cerebral cortex layers are all annotated as glutamatergic neurons.

Transferring cell type annotations to corresponding snATAC-seq

Cell type annotations were added to each ArchR project using the per-cell metadata extracted from Seurat objects. Barcode sequences were matched between assays and annotations carried over from snRNA-seq analysis with no modifications.

Differential gene expression analysis of pseudobulk snRNA-seq

The raw, unnormalized counts were extracted from the annotated Seurat object for subtypes of interest and summed across all nuclei in each individual mouse for a sample-level pseudobulk counts matrix across all expressed genes. Using pydeseq2[265], defined groups such as sex were compared within subtypes. Results were filtered by an absolute log fold change >1 and adjusted p-value < 0.01.

Pseudotime ordering of dynamic cell states in hippocampus

Cell types of interest were subset from the tissue-level Seurat object for pseudotime ordering using Monocle 3[100,266–269]. The root cells were chosen for "order_cells" according to the known stage of the cells. The oligodendrocytes and OPCs were subset from the hippocampus dataset, with root cells corresponding to the OPCs. For ordering of the DG cells, root cells correspond to the cells from early timepoints. Pseudotime values for the ordered cells were incorporated into their metadata for downstream analysis.

Calculating single-nucleus regulatory topics using Topyfic

The raw, unnormalized counts were extracted from each filtered Seurat object per tissue and barcoding technology (Parse and 10x). Genes were filtered to 2,701 regulatory genes[128] determined by microRNA-host gene correlations, annotated transcription factors, and genes annotated with the following Gene Ontology (GO) terms: 0004402 (histone acetyltransferase activity), 0004407 (histone deacetylase activity), 0042054 (histone methyltransferase activity), 0032452 (histone demethylase activity), 0016592 (mediator complex), 0006352 (DNA-templated transcription, initiation), 0003682 (chromatin binding), 0006325 (chromatin organization), 0030527 (structural constituent of chromatin), and 0140110 (transcription regulator activity). MicroRNA host genes were included if they are annotated as a host gene (e.g. *Mir133a-1hg*, *Mir124a-1hg*) and/or their Spearman correlation with expression of the mature microRNA was ≥ 0.3[128].

Depth normalization was performed on each raw counts matrix by tissue (x 5) and technology (Parse and 10x; 10 total matrices) by a round of proportional fitting followed by log transformation, then another round of additional proportional fitting[270]. An anndata object was constructed from the normalized matrix, 2,701 regulatory genes, and per-cell metadata including subtype and celltype annotations.

Topyfic was run with a range of k values for each tissue and technology using 100 runs of LDA with batch_size of 128 and 5 minimum iterations[128]. The best k per tissue and technology was determined by comparing k to the number of resulting topics, n. The closest k to the resulting n value was chosen: $k = 15$ for Parse and 13 for 10x adrenal, 14 for Parse and 13 for 10x cortex, 13 for Parse and 21 for 10x hippocampus, 11 for Parse and 13 for 10x heart, and 12 for Parse and 8 for 10x gastrocnemius. Harmony[227] was used to combine the best models learned separately from each technology to a unified set of topics, filtering out topics with participation in less than 1% of nuclei in the smaller of the two datasets. Downstream analysis such as comparisons between topics was facilitated by analysis of the gene weights in each topic (Tables S4-S8).

Topics analysis

Harmonized snRNA-seq topics in each tissue were characterized by analysis of topic-trait enrichment (Topyfic function "TopicTraitRelationshipHeatmap" on the analysis TopModel object), a measurement of how highly-weighted topic genes are specifically expressed in traits like celltypes, subtypes, ages, and sexes[128]. Topics were further interpreted by cell participation across celltypes and subtypes, represented as pie charts (function "pie_structure_Chart") and structure plots (function "structure_plot")[128]. Two specific topics of interest, such as immune-related topics in heart and brain, were compared using an MA plot (function "MA_plot"), and topics were compared across tissues by Pearson correlation based on gene weights[128].

Characterizing ENCODE cCRE specificity with snATAC-seq

The ENCODE V4 catalog of candidate cis-regulatory elements (cCREs) for mm10 was downloaded from the ENCODE portal (`https://www.encodeproject.org/files/ENCFF167FJQ/`)[248].

All 926,843 cCREs were added to each tissue's ArchR project by the function "addPeakSet", then scored using "addPeakMatrix", which counts the number of fragments per region with a maximum count of 4 to prevent large biases in the counts[249]. The raw counts matrices were extracted ("getMatrixFromProject"), pseudobulked by integrated snRNA cluster, and normalized by RPM. RPKM was not used due to the limited distribution of cCRE lengths, between 150 and 350 bp with a mean of 269 bp and standard deviation of 64.9 bp (Fig. 3.15). For clarity in downstream analysis, small clusters of less than 100 multiome nuclei were removed (such as a cluster corresponding to 16 hepatocytes detected in adrenal gland, most likely a dissection artifact). Each cCRE was classified as accessible in a celltype if it scored ≥ 5 RPM in at least one cluster corresponding to that celltype. Categories of "specific", "shared", "general", or "global" were assigned based on the number of celltypes within and across tissues with open chromatin at each cCRE. "Specific" refers to cCREs accessible in only one celltype above the RPM threshold across all tissues. Common celltypes such as macrophages and endothelial cells were considered one celltype. "Shared" refers to cCREs accessible in more than one celltype within or across tissues. "General" refers to cCREs accessible in all major celltypes within a tissue, and "global" refers to cCREs accessible in all major celltype across all tissues. Major celltypes were defined as those whose cumulative sum makes up 90% of the cell types in the tissue; for example neurons in the brain, myonuclei in skeletal muscle, and adrenal cortical cells, followed by other major types such as glial cells, endothelial cells, and fibroblasts.

Differential accessibility analysis of pseudobulk snATAC-seq

Pseudobulk cCRE counts matrices were generated per sample and tissue by extracting raw single-nucleus counts and summing per cCRE across all nuclei from each individual mouse. Using pydeseq2[265], accessibility of the previously characterized cCREs accessible in pseudobulk clusters was compared between sexes and timepoints within each tissue and group,

i.e. female vs. male adrenals at PND 14, female vs. male adrenals at 2 months, PND 14 vs. 2 month male adrenals, PND 14 vs. 2 month female adrenals, etc. Results were filtered by an absolute log fold change >2 and adjusted p-value < 0.01. Unique cCREs open in each group were counted and normalized by the total number of cCREs accessible in the tissue.

Motif enrichment analysis

Motif enrichment was calculated using ArchR to analyze transcription factor activity in celltype specific cCREs. The JASPAR2024 CORE vertebrate non-redundant PFMs[271] were formatted as a custom RangedSummarizedExperiment, and matches with the full set of cCREs were extracted with motifmatchr[272,273]. ArchR's "customEnrichment" function was used to run hypergeometric-based enrichment testing on the matched motifs and a custom subset of specific cCREs as a GenomicRanges object[249,273]. Motifs were filtered by bulk RNA-seq expression in each tissue for downstream analysis (>5 TPM in at least 1 sample).

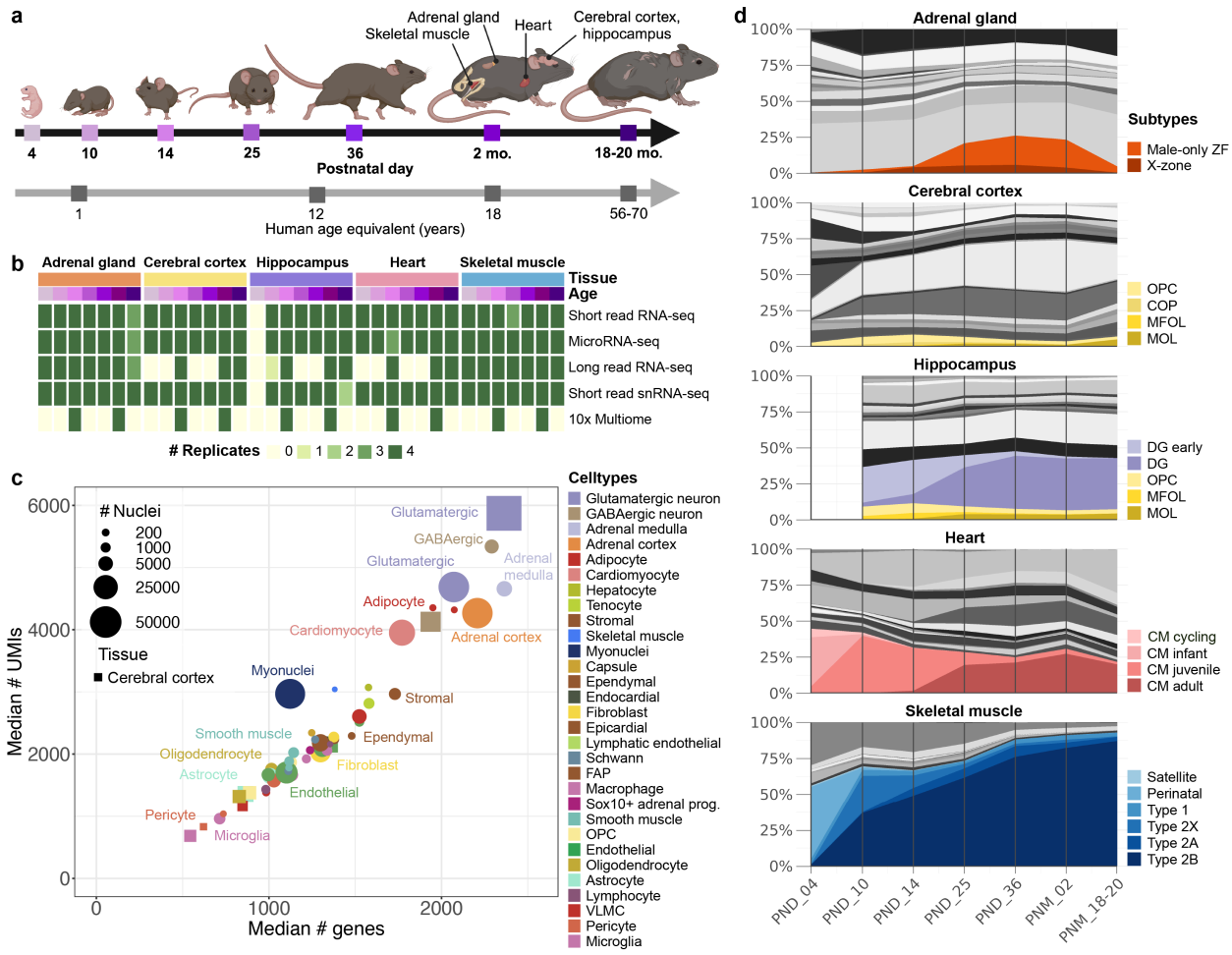

Figure 3.1: **Overview of the ENCODE4 mouse dataset of postnatal development.** **a,** Samples from 5 coordinated B6/CAST F1 hybrid mouse tissues were collected at 7 key timepoints from postnatal day 4 to 18-20 months (excluding hippocampus, which was collected from PND 10 onwards). The mouse postnatal timecourse corresponds to human infancy to late adulthood. **b,** Overview of the sampled tissues, timepoints, and assays from each tissue in the ENCODE mouse dataset. The majority of assays have successful experiments in 4 replicates, 2 males and 2 females, per timepoint. 10x Multiome experiments were selectively performed on PND 14 and 2 month timepoints. **c,** Comparison of gene and UMI counts in cell types across all five tissues, with point sizes reflecting the number of nuclei in each cell type within its respective tissue. In common brain cell types, cerebral cortex data points are represented by squares. **d,** Dynamics of cell subtype composition across postnatal development in all five tissues. Highlighted subtypes are shown in color, while all others are represented in shades of grey (see Fig. 3.5, 3.6, 3.7, 3.8, 3.9 for full-color versions).

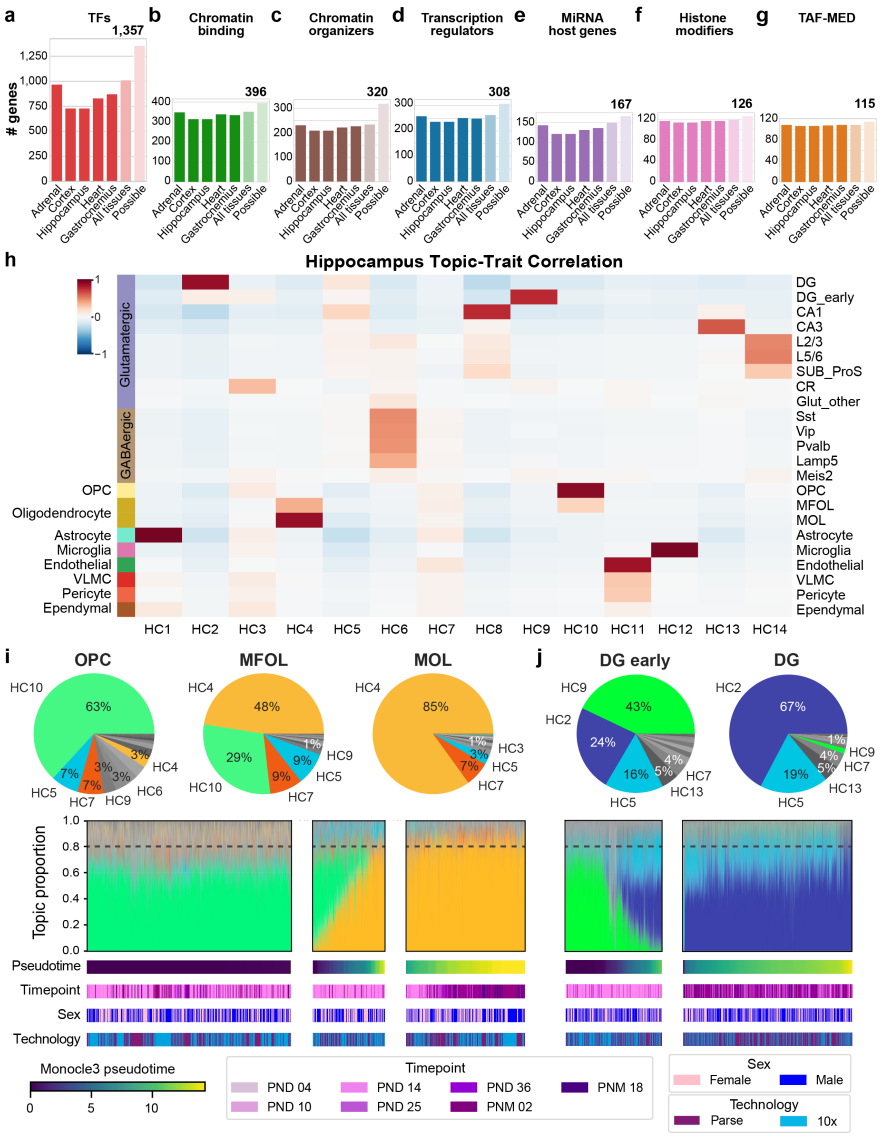

Figure 3.2: **Characterization of hippocampus topics in annotated subtypes. a,** Number of transcription factors detected at > 1 TPM in bulk RNA-seq data in each tissue. Sixth column reports the union of TFs in all tissues, and the last column reports the total number of TFs in our regulatory gene set. **b,** Number of chromatin binding genes detected in bulk RNA-seq data. **c,** Number of chromatin organizing genes detected in bulk RNA-seq data. **d,** Number of transcription regulators detected in bulk RNA-seq data. **e,** Number of host genes representing microRNAs. **f,** Number of histone modifying genes such as acetyltransferases, deacetylases, methyltransferases, and demethylases detected in bulk RNA-seq data. **g,** Number of TBP-associated factors and members of the Mediator complex detected in bulk RNA-seq data. **h,** Topic-trait relationship heatmap between 14 hippocampus topics and 10 cell types (23 subtypes). **i,** Proportion of topics in OPC (oligodendrocyte precursor), MFOL (myelin-forming oligodendrocyte), and MOL (mature oligodendrocyte) subtypes summarized in pie charts and displayed as a compressed stacked bar plot (structure plots) for single nuclei ordered by pseudotime. Pseudotime, timepoint, sex, and snRNA-seq barcoding technology are indicated for each nucleus below the structure plots. **j,** Proportion of topics in early DG (dentate gyrus) and DG subtypes.

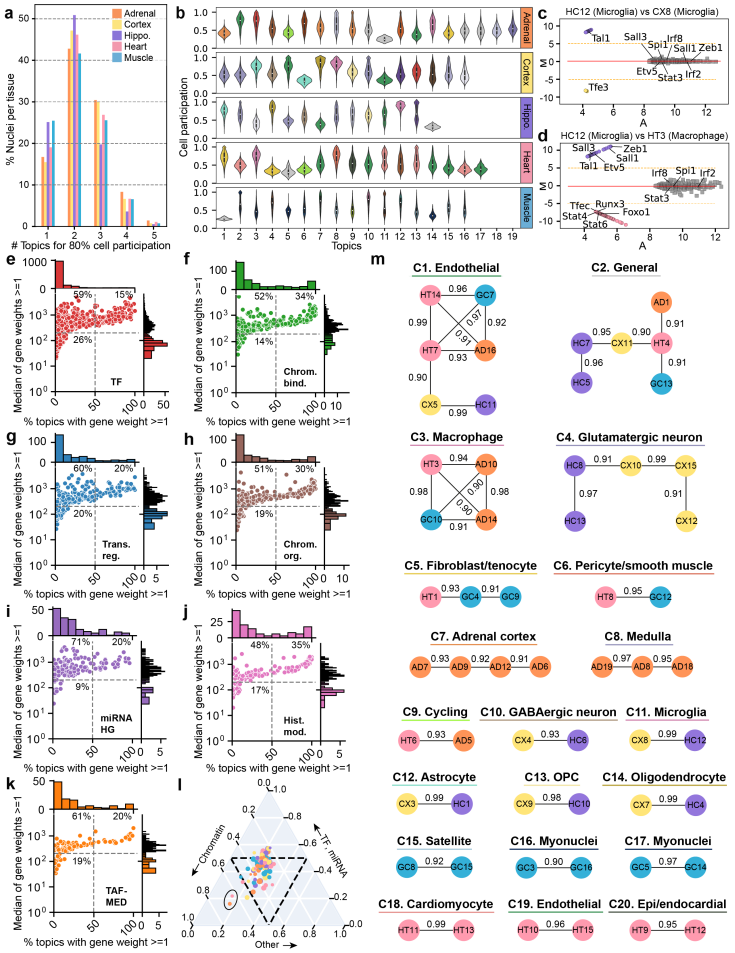

Figure 3.3: **Characterization of topics across diverse tissues. a,** Comparison of the number of topics required to constitute 80% of cell participation when sorted from the largest to the smallest proportion per nucleus, along with the percentage of nuclei in each category out of the total nuclei per tissue. **b,** Distribution of cell participation in each topic across all five tissues, with violins colored by associated celltype, when possible (see Fig. 3.1c for color legend). **c,** MA plot comparing microglia-specific HC12 with microglia-specific CX8. X-axis (A) represents average weight of the gene between both topics in the comparison, and y-axis (M) represents log base 2 of the fold change of gene weight between topics. Genes of interest are labeled. **d,** MA plot comparing HC12 with macrophage-specific HT3. **e,** Percent of topics containing each gene in the TF biotype vs. median of the gene's weight across all topics when the gene weight is >= 1. Percentages of genes in each quadrant, out of the total number in the biotype, are labeled. Percent of topics containing each gene in each biotype vs. median weight across topics for **f,** chromatin binders, **h,** chromatin regulators, **i,** microRNA host genes, **j,** histone modifiers, and **k,** TAF-MED complex-associated genes. **l,** Gene biotype simplex with a sector for chromatin (left), encompassing chromatin binders, chromatin regulators, and histone modifiers, a sector for TFs and microRNA host genes (top), and a sector for all other biotypes (right). Topics are color-coded by tissue and scaled by number of genes. **m,** 20 clusters of correlated topics (C1 - C20), filtered to connections >= 0.9 cosine similarity. Each node represents a topic, color-coded by tissue, and edges labeled by cosine similarity score calculated on the basis of gene weights between topics.

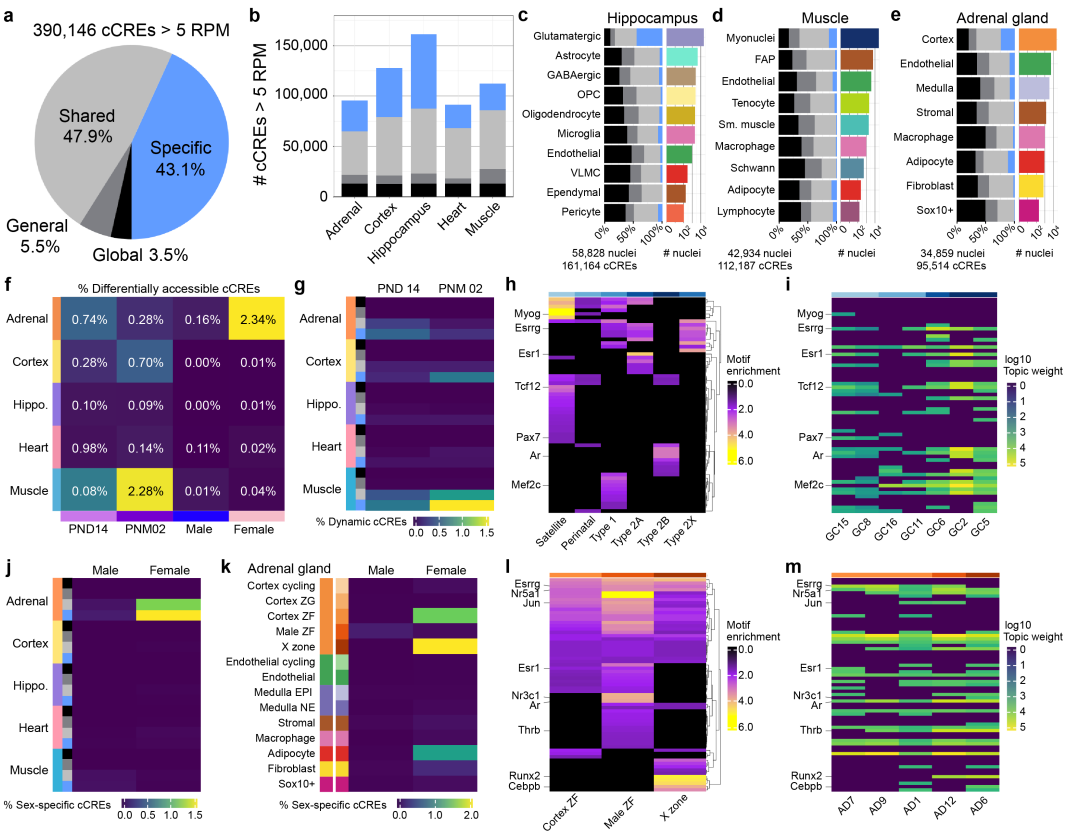

Figure 3.4: **Characterization of celltype-specific candidate cis-regulatory elements and motif enrichment analysis. a,** 390,146 ENCODE mm10 cCREs filtered by > 5 RPM in 10x snATAC-seq data pseudobulked by integrated snRNA-seq clusters. Specific cCREs in blue (168,443) are accessible in only one celltype above 5 RPM across all tissues, shared in grey (186,805) are accessible in more than one celltype within or across tissues, general in dark grey (21,314) are accessible in all major celltypes within a tissue, and global in black (13,584) are accessible in all major celltype across all tissues. **b,** Number of cCRE per specificity category in each tissue. **c,** Breakdown of cCRE specificity by percent of cCREs detected in each celltype in hippocampus and number of nuclei per celltype in 10x Multiome. Breakdown of cCRE specificity and total number of nuclei per celltype in **d,** gastrocnemius and **e,** adrenal gland. **f,** Percentage of the cCREs detected in each tissue with significant increase in accessibility in each group compared to its counterpart across all tissues. **g,** Overlap of differentially accessible cCREs between timepoints with specificity categories, reported as percent differentially accessible out of total detected in each tissue. **h,** Motif enrichment (adj. p-value < 0.05) of expressed TFs (TPM > 5 in at least 1 bulk RNA-seq sample) in satellite, perinatal, and myonuclear fiber type-specific cCREs. **i,** Weight of TFs as ordered in h across topics corresponding to satellite and myonuclear subtypes. **j,** Overlap of differentially accessible cCREs between sexes with specificity categories, reported as percent differentially accessible out of total detected in each tissue. **k,** Overlap of sex-specific cCREs with celltype-specific cCREs, reported as percent differentially accessible out of total detected in each tissue. **l,** Motif enrichment (adj. p-value < 0.05) of expressed TFs (TPM > 5 in at least 1 bulk RNA-seq sample) in adrenal ZF subtype-specific cCREs. **m,** Weight of TFs as ordered in l across topics corresponding to adrenal ZF subtypes.

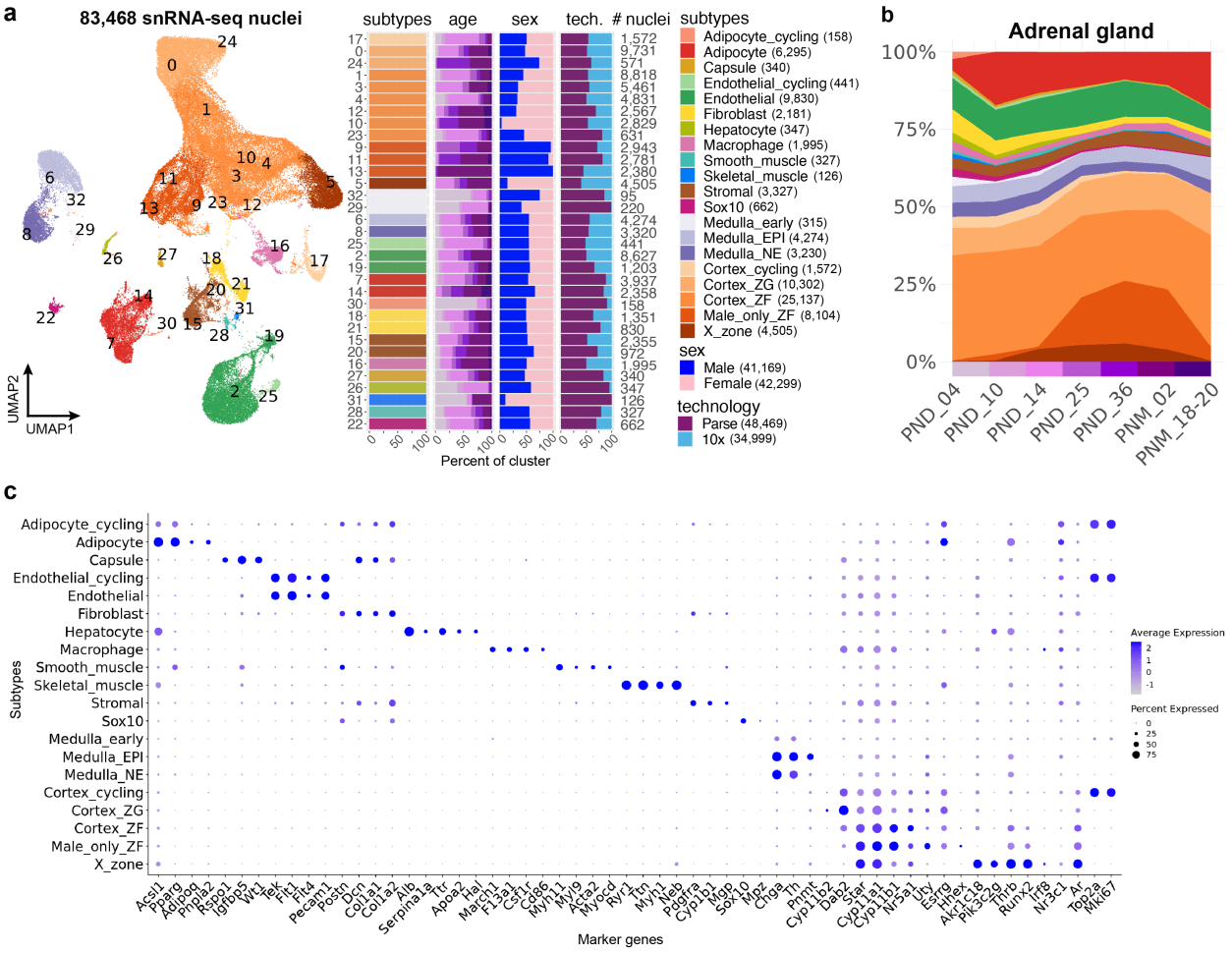

Figure 3.5: **Clustering and annotation of integrated adrenal gland snRNA-seq data.** a, UMAP representation of 83,468 adrenal gland nuclei integrated between Parse and 10x Multiome platforms and breakdown of age, sex, and technology per cluster. Numbers of nuclei per cluster are annotated to the right of the bar plots, and numbers of nuclei per annotated cell subtype are included in the legend. b, Dynamics of cell subtype composition across postnatal development in adrenal gland, with the same color legend as in a. For consistent sampling at each timepoint, only Parse data is shown. c, Expression of marker genes across subtypes in adrenal gland.

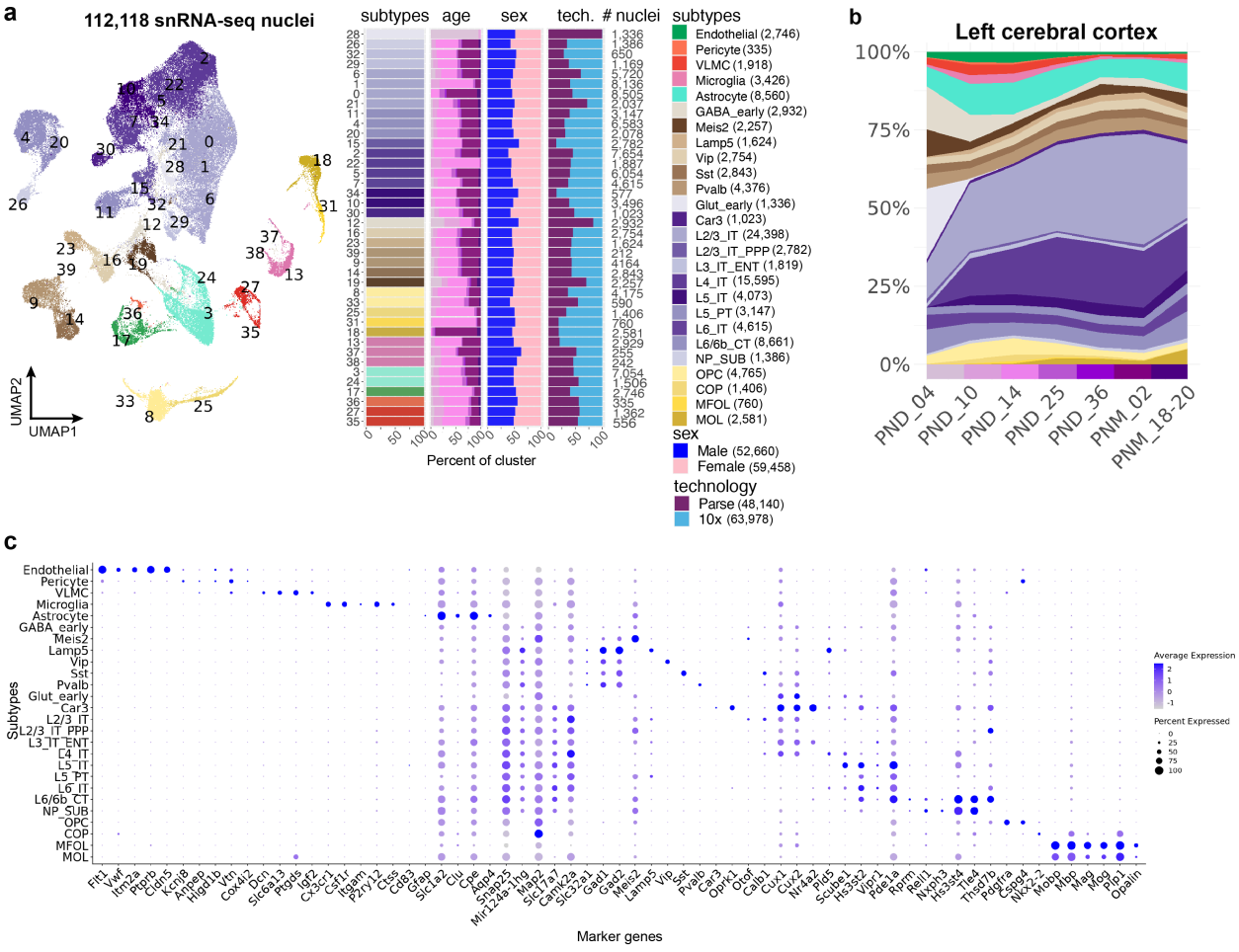

Figure 3.6: **Clustering and annotation of integrated left cerebral cortex snRNA-seq data.** a, UMAP representation of 112,118 left cerebral cortex nuclei integrated between Parse and 10x Multiome platforms and breakdown of age, sex, and technology per cluster. Numbers of nuclei per cluster are annotated to the right of the bar plots, and numbers of nuclei per annotated cell subtype are included in the legend. b, Dynamics of cell subtype composition across postnatal development in cortex, with the same color legend as in a. For consistent sampling at each timepoint, only Parse data is shown. c, Expression of marker genes across subtypes in cortex.

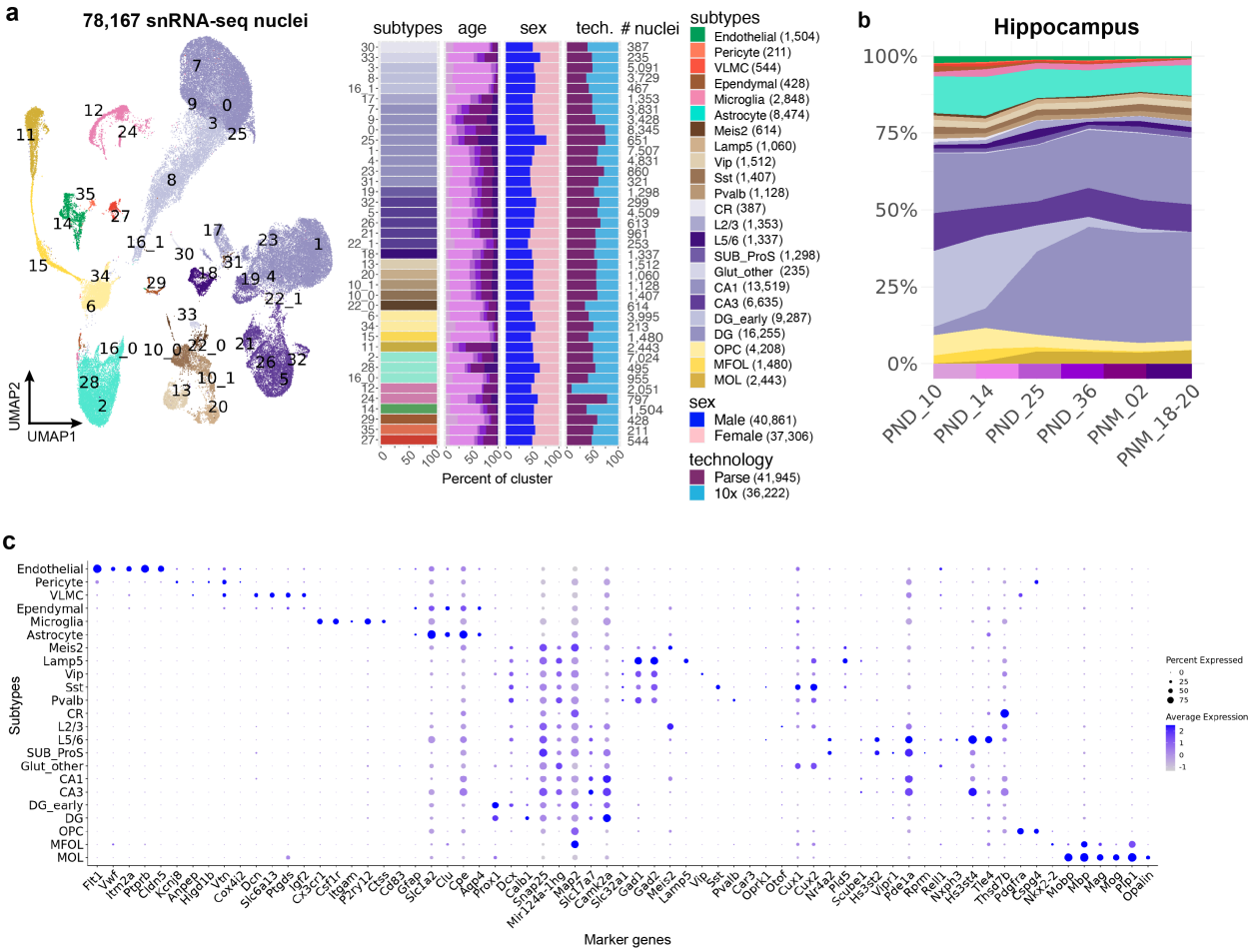

Figure 3.7: **Clustering and annotation of integrated hippocampus snRNA-seq data.** a, UMAP representation of 78,167 hippocampus nuclei integrated between Parse and 10x Multiome platforms and breakdown of age, sex, and technology per cluster. Numbers of nuclei per cluster are annotated to the right of the bar plots, and numbers of nuclei per annotated cell subtype are included in the legend. b, Dynamics of cell subtype composition across postnatal development in hippocampus, with the same color legend as in a. For consistent sampling at each timepoint, only Parse data is shown. c, Expression of marker genes across subtypes in hippocampus.

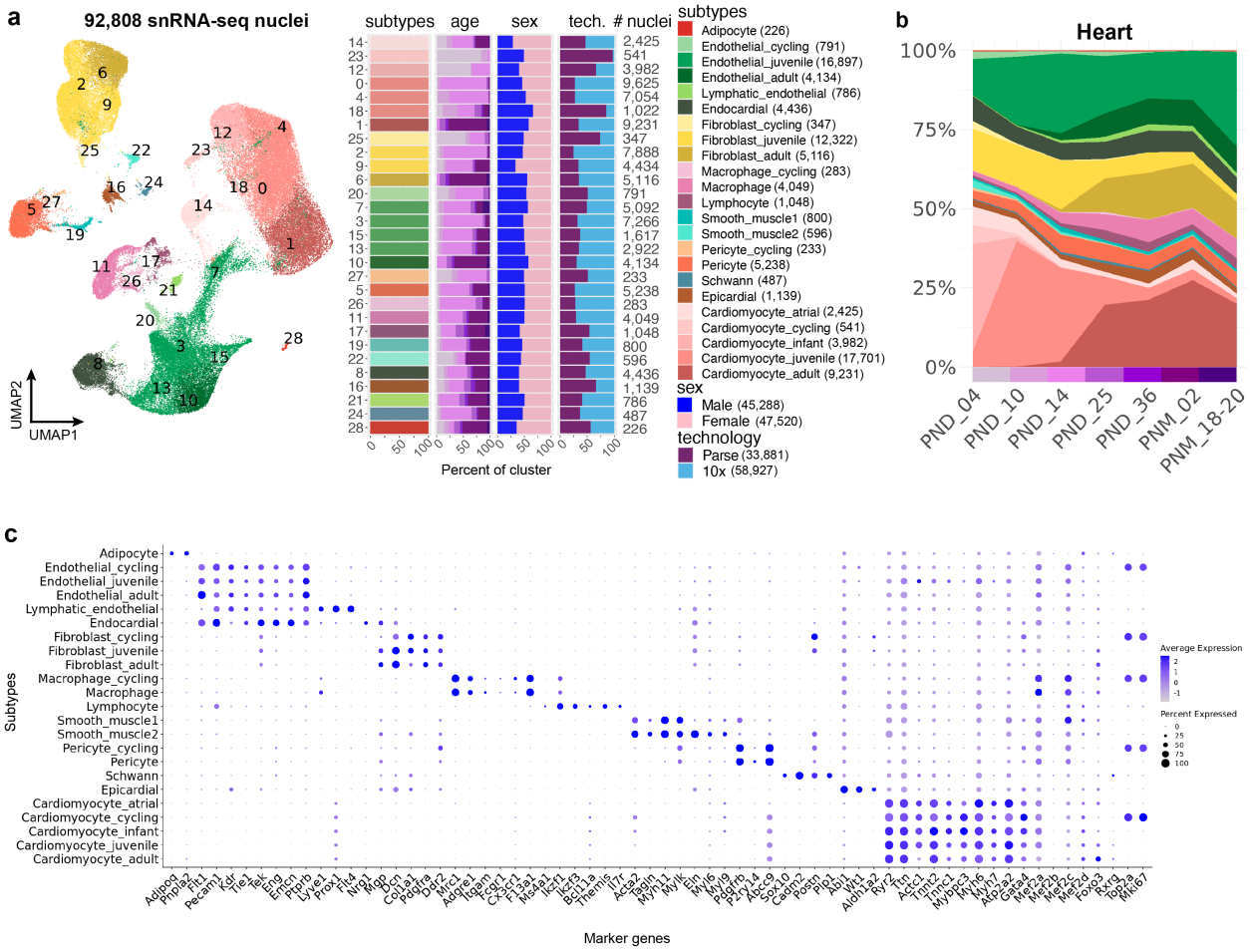

Figure 3.8: **Clustering and annotation of integrated heart snRNA-seq data.** a, UMAP representation of 78,167 heart nuclei integrated between Parse and 10x Multiome platforms and breakdown of age, sex, and technology per cluster. Numbers of nuclei per cluster are annotated to the right of the bar plots, and numbers of nuclei per annotated cell subtype are included in the legend. b, Dynamics of cell subtype composition across postnatal development in heart, with the same color legend as in a. For consistent sampling at each timepoint, only Parse data is shown. c, Expression of marker genes across subtypes in heart.

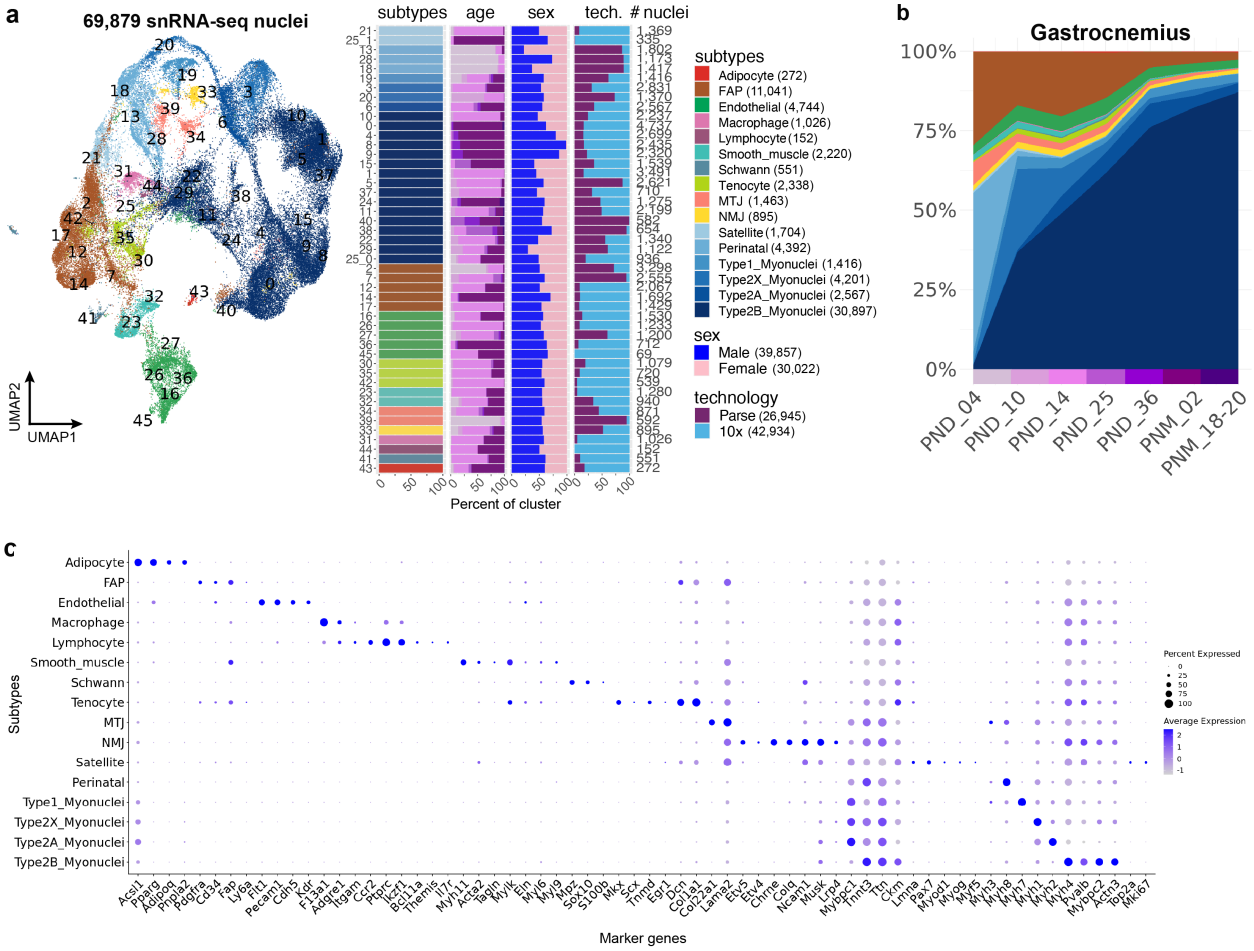

Figure 3.9: **Clustering and annotation of integrated gastrocnemius snRNA-seq data.** a, UMAP representation of 69,879 gastrocnemius nuclei integrated between Parse and 10x Multiome platforms and breakdown of age, sex, and technology per cluster. Numbers of nuclei per cluster are annotated to the right of the bar plots, and numbers of nuclei per annotated cell subtype are included in the legend. b, Dynamics of cell subtype composition across postnatal development in gastrocnemius, with the same color legend as in a. For consistent sampling at each timepoint, only Parse data is shown. c, Expression of marker genes across subtypes in gastrocnemius.

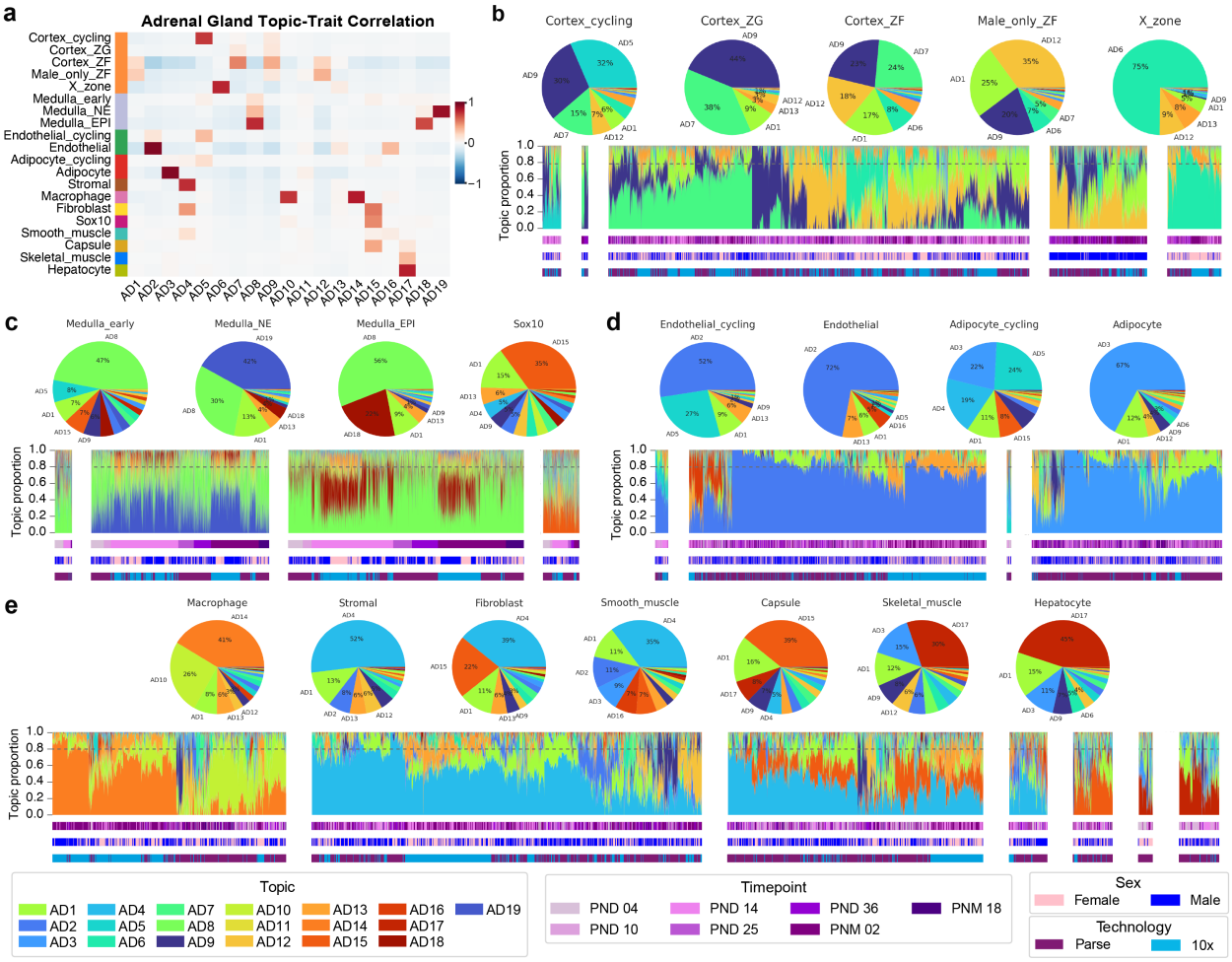

Figure 3.10: **Regulatory topic enrichment and proportions in adrenal gland cell subtypes.** a, Topic-trait correlation in 19 regulatory adrenal topics. b, Structure plots in adrenal cell subtypes, summarized in above pie charts. Topics AD7, AD9, AD12, and AD6 are specific to adrenal cortex. c, AD19, AD8, and AD18 are specific to adrenal medulla, while AD15 is specific to *Sox10*+ progenitor cells. d, AD2 is endothelial-specific and AD3 is adipocyte-specific. AD5 is a general cycling topic enriched in proliferating cells regardless of subtype. e, Topics AD14 and AD10 are specific to macrophages, and topic AD4 is shared across stromal, fibroblast, and smooth muscle cells. AD15 is enriched in the adrenal capsule and fibroblasts.

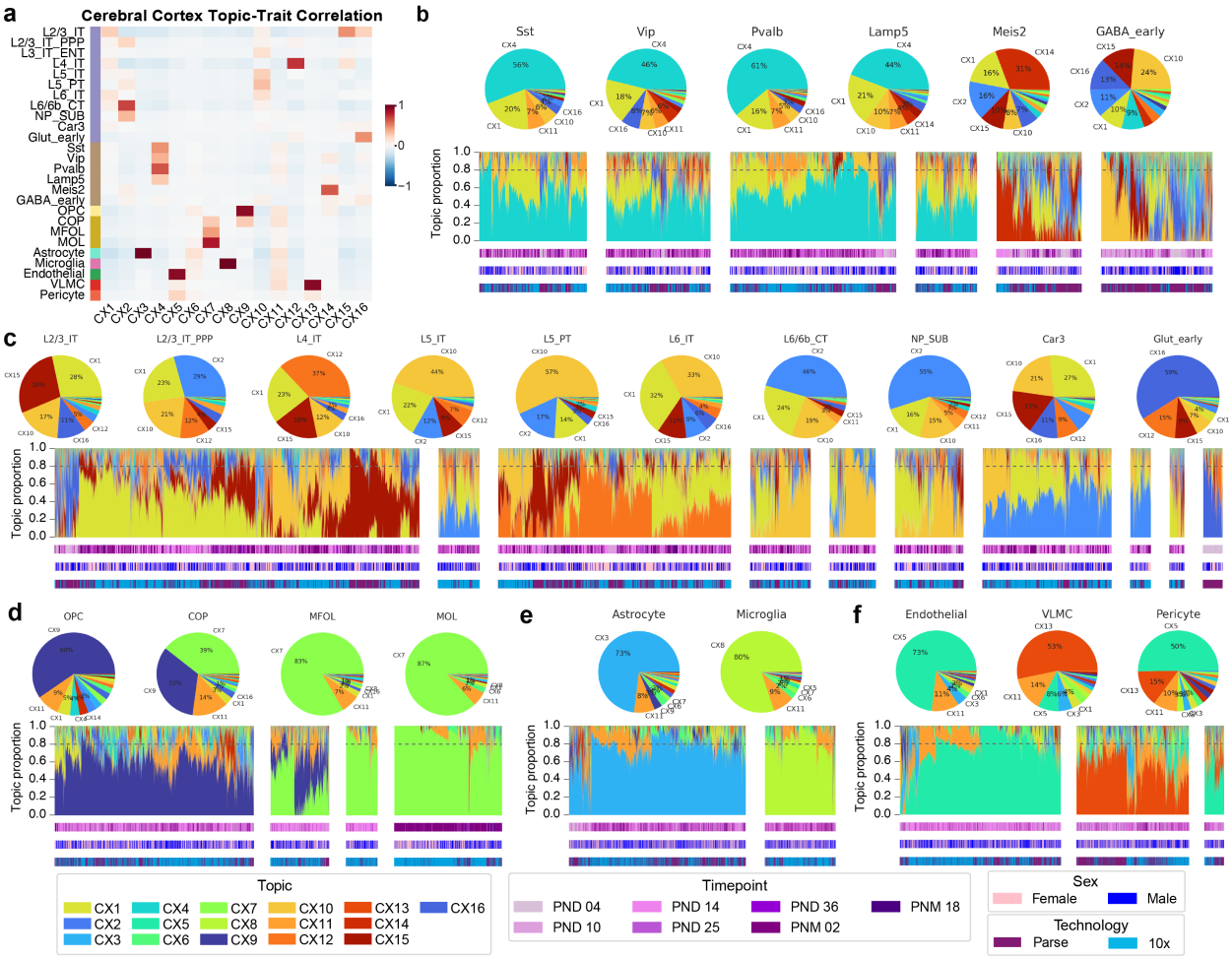

Figure 3.11: **Regulatory topic enrichment and proportions in left cerebral cortex cell subtypes.** a, Topic-trait correlation in 16 regulatory cortex topics. b, Structure plots in cortex cell subtypes, summarized in above pie charts. CX4 is a general GABAergic topic other than *Meis2+* and early GABAergic cells, which are described by a mix of topics. c, Topics CX1, CX2, CX10, and CX12 are all enriched in various excitatory neuronal subtypes. d, CX9 is enriched in OPC and COP progenitors, while CX7 is enriched in mature oligodendrocytes. e, CX3 is astrocyte-specific and CX8 is microglia-specific. f, CX5 is enriched in endothelial and pericytes and CX13 is specific to VLMC (vascular leptomeningeal cells).

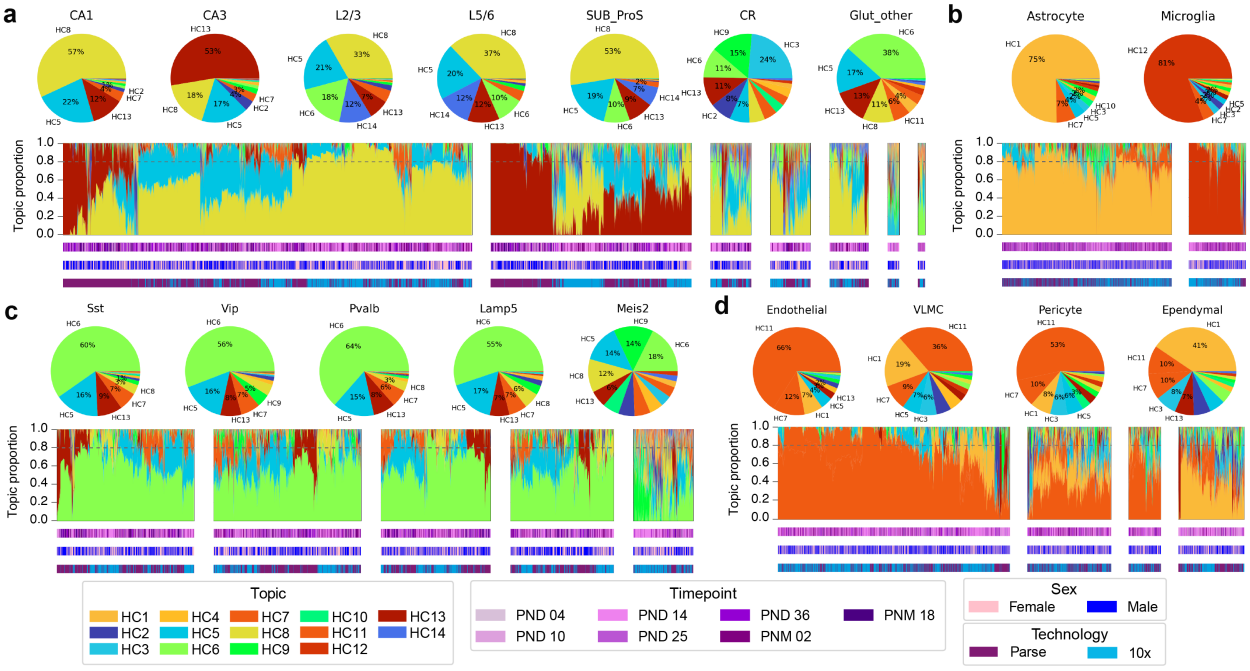

Figure 3.12: **Regulatory topic proportions in left hippocampus cell subtypes.** a, Structure plots in hippocampus cell subtypes, summarized in above pie charts. HC8 is enriched in CA1 and shared across various other glutamatergic subtypes, and HC13 is CA3-specific. b, HC1 is astrocyte-specific, while HC12 is microglia-specific. c, HC6 and HC5 are general GABAergic neuron topics, while the *Meis2+* subtype is described by a mix of topics. d, HC11 is enriched in endothelial, pericytes, and VLMC (vascular leptomeningeal cells), while HC1 is shared in VLMC and ependymal cells.

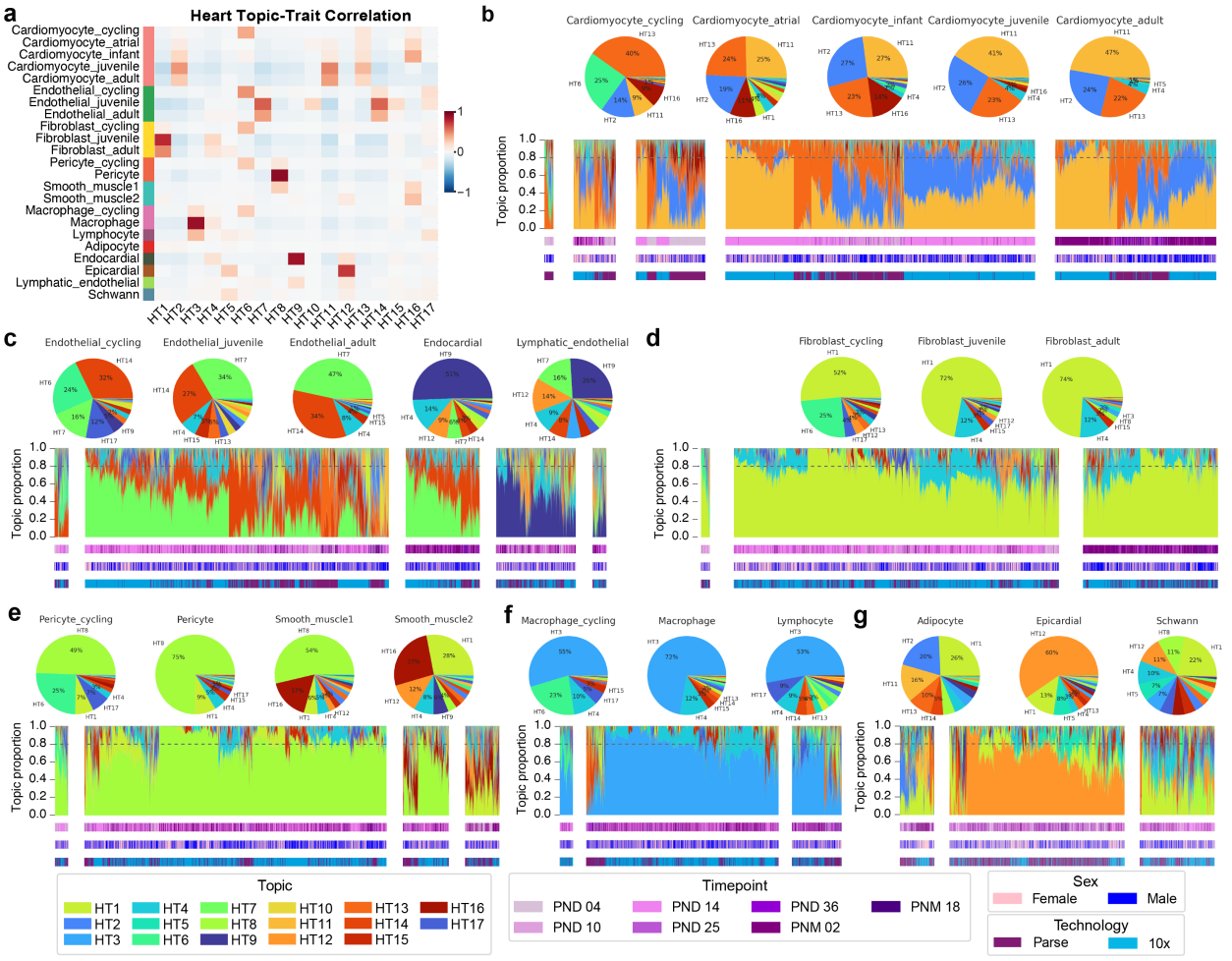

Figure 3.13: **Regulatory topic enrichment and proportions in heart subtypes.** a, Topic-trait correlation in 17 regulatory heart topics. b, Structure plots in heart cell subtypes, summarized in above pie charts. Topics HT2, HT11, and HT13 are shared by cardiomyocytes at all developmental stages. c, HT7 and HT14 are enriched in endothelial cells, while HT9 is enriched in endocardial and lymphatic endothelial cells. d, HT1 is enriched in cardiac fibroblasts at all developmental stages. e, HT8 is specific to pericytes and one subtype of smooth muscle, while the other smooth muscle subtype is enriched in HT16 and HT1. f, HT3 is the macrophage-specific topic in heart. HT6 is a general cycling topic enriched in proliferating cells regardless of subtype. g, HT12 is specific to epicardial cells. Adipocytes and Schwann cells are made up of several topics, the largest fraction being HT1 which is also shared with fibroblasts and smooth muscle.

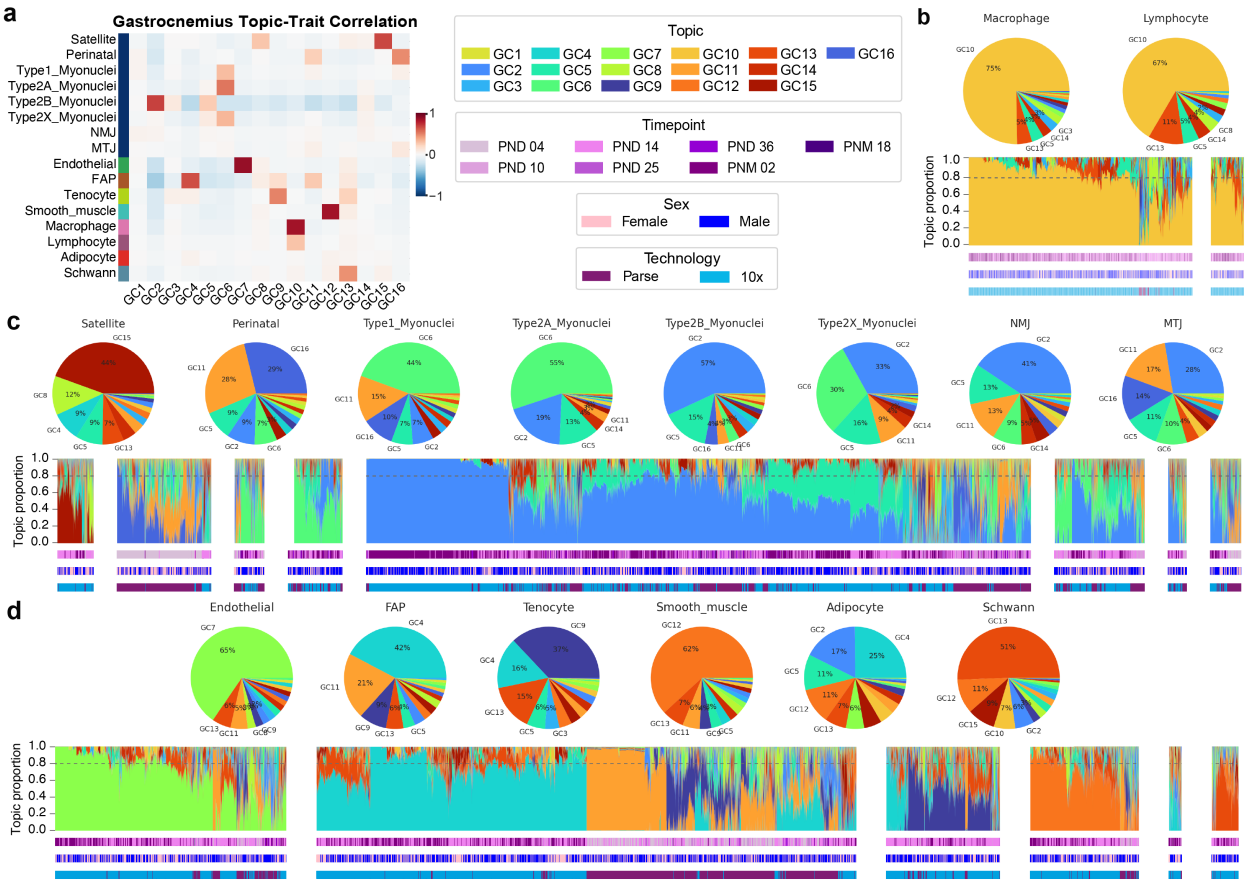

Figure 3.14: **Regulatory topic enrichment and proportions in gastrocnemius subtypes.** a, Topic-trait correlation in 16 regulatory gastrocnemius topics. b, Structure plots in gastrocnemius cell subtypes, summarized in above pie charts. GC10 is enriched in both macrophages and lymphocytes. c, Topics GC2, GC5, GC6, and GC11 are shared across mature myofiber subtypes. Most cell participation in type 2B and type 2X is attributed to topic GC2, but type 2X also shares GC6 with type 2A and type 1. Perinatal myonuclei are described by GC16 and GC11, while GC15 and GC8 are specific to satellite cells. Specialized NMJ (neuromuscular junction) and MTJ (myotendinous junction) myonuclei have no specific regulatory topic, but share a mix of muscle-enriched topics. d, GC7, GC12, and GC13 are specific to endothelial, smooth muscle, and Schwann subtypes, respectively. FAP (fibro-adipogenic progenitors) are enriched for GC4 and GC11 which are also timepoint-specific, with GC11 enriched in infants and GC4 specific to adults and juveniles.

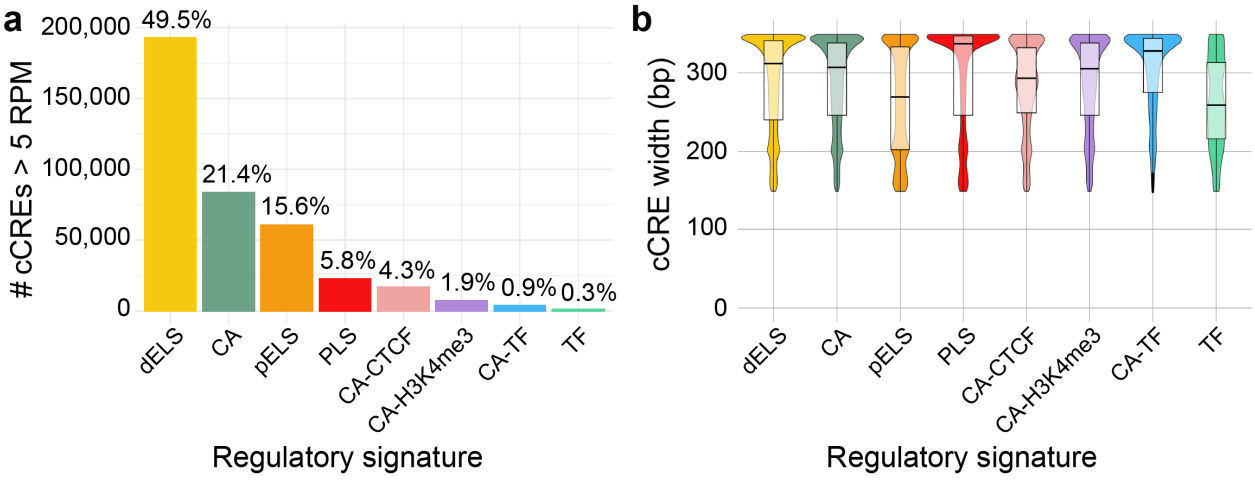

Figure 3.15: **cCRE classification by regulatory signature.** a, Breakdown of 390,146 cCREs >5 RPM in at least 1 pseudobulk cluster in 10x Multiome snATAC-seq data across all 5 tissues. Most cCREs are classified as dELS (distal enhancer-like signature), CA (chromatin accessible), and pELS (proximal enhancer-like signature). Less than 15% of accessible cCREs are CA-CTCF (chromatin-accessible CTCF), CA-H3K4me3 (chromatin-accessible with promoter-associated histone modification), CA-TF (chromatin-accessible, TF signal), and TF (TF signal). b, All cCREs are between 150 and 350 bp with an average of 284 bp with consistent distributions across the 8 categories. Therefore, we opted to normalize snATAC-seq quantifications across the cCREs using reads-per-million (RPM).

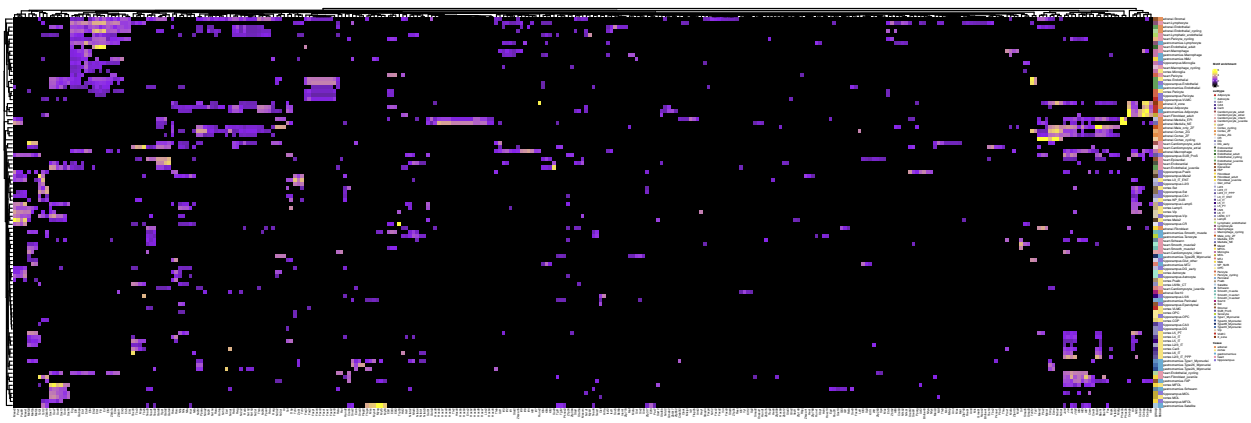

Figure 3.16: **Motif enrichment in subtype-specific cCREs across all tissues.** Out of 765 possible JASPAR motifs, 317 were enriched in at least 1 subtype with an adjusted p-value ≤ 0.05, enrichment ≥ 1.5, and bulk RNA-seq expression ≥ 5 TPM in at least 1 sample in the tissue.

Chapter 4

Characterizing the impact of genetic diversity on gene expression across adult cell types and states in mice

4.1 Abstract

Mapping the impact of genomic variation on gene expression is crucial for understanding the molecular basis of complex phenotypic traits. The identification of genomic loci quantitatively associated with regulation of gene expression is a major focus of genetics but is affected by factors such as the large sample size required and the significant influence of environmental factors on biological traits. Mouse models of natural genetic diversity overcome these problems, providing a controlled and reproducible framework for capturing the breadth of genomic variation observed in different genotypes. As part of the IGVF consortium's efforts to catalog the effects of genetic variation on tissues, we characterize the transcriptional landscape of the mouse founder strains for the Collaborative Cross and Diversity Outbred that com-

prise 5 classical inbred strains (C57BL6/J, A/J, NOD/ShiLtj, NZO/HlLtJ, 129S1/SvImJ) and 3 wild-derived strains (PWK/PhJ, WSB/EiJ, and CAST/EiJ). We sequenced samples from 8 tissue groups for 4 male and 4 female replicates per mouse strain using single-nucleus RNA-seq to generate an "8-cube" dataset of 5.9 million nuclei passing our filters across 108 cell types and cell states. We observe that genetic divergence correlates with transcriptional variation across most cell types, with most transcriptional variation found in PWK and Castaneus. Further analysis of specific cell types revealed substantial variation in common laboratory strains associated with known traits in those strains. The characterization of these founders will enable the interpretation of gene expression in matching tissues from F1 hybrids and Collaborative Cross lines as well as facilitate the identification of cis and trans contributions to gene expression variation and eQTL mapping at the cell type level. The founder 8-cube dataset presented here lays the foundation for advancing our systematic understanding of the genomic basis for cell type-specific transcriptional regulation.

4.2 Introduction

Understanding the impact of genetic variation on gene expression is fundamental for deciphering the molecular mechanisms underlying complex traits, diseases, and developmental processes. Genetic variation, such as single nucleotide polymorphisms (SNPs) and structural variants, can modulate gene expression levels by affecting regulatory elements such as promoters, enhancers, and transcription factor binding sites. This modulation can lead to phenotypic diversity among individuals, including susceptibility to diseases, response to therapies, and variation in physiological traits. Studying the influence of genetic variants on gene expression across different tissues, disease states, and developmental stages provides crucial insights into the regulatory networks governing cellular functions. Cell-type-specific gene expression patterns are essential for the proper function of tissues and systems within

the body. Researchers can unravel cell-type-specific regulatory mechanisms and identify potential targets for therapeutic interventions by examining how genetic variants impact gene expression in certain cell types, A central aim of the IGVF (Impact of Genomic Variation on Function) consortium is to comprehensively map the influence of genetic variants on molecular traits such as gene expression across diverse tissues and disease contexts in human and mouse[18].

Recombinant inbred (RI) mouse strains are a valuable platform for investigating complex mammalian traits and diseases[24,59]. They are generated by crossing genetically diverse founder strains to produce F1 hybrids. These F1 hybrids are then bred together to generate a large set of F2s, which are paired as founders for strains inbred over 20 generations[24]. This results in RI strains with stabilized genetic compositions that are genetically identical within each strain but differ between strains, ideal for genetic mapping studies[24]. More recently, numerous diverse strains have been crossed with the specific aim of analyzing complex traits. This effort has resulted in the creation of widely used community resources like the Collaborative Cross (CC) RI panel and the Diversity Outbred (DO) panel[58,62]. The CC panel is an invaluable tool for modeling natural genetic diversity while also retaining the benefits of inbred characteristics such as stable, homozygous genomes, ensuring reproducibility and consistency across experiments[58,64]. Each CC strain represents a unique combination of haplotypes from 8 founder strains, comprising five laboratory strains (C57BL/6J, NOD/ShiLtJ, NZO/HlLtJ, A/J, 129S1/SvImJ) and three wild-derived strains (WSB/J, PWK/PhJ, and CAST/EiJ). While the DO panel is based on the same 8 founders, random crosses result in unique heterozygous genomes that more closely mirror the diversity observed in human genomes and refine the landscape of QTL mapping[62]. The 8 founder strains collectively encompass a broad spectrum of natural genetic variation that enable investigation of phenotypes within a controlled yet genetically diverse framework[60–62,64].

While C57BL/6J ("B6") is the most used laboratory strain, other strains are also used

for specific disease research. For example, NOD/ShiLtJ (non-obese diabetic, "NOD") is a commonly used model for type 1 diabetes and NZO/HlLtJ (New Zealand obese, "NZO") for type 2 diabetes[32,33]. A/J ("AJ") also serves as a model for asthma[35], emphysema[36], and age-onset muscular dystrophy[37]. Other lab strains are preferred for specific experimental techniques, such as embryonic stem cell (ESC) derivation. Although ESC lines are now available for all 8 founder strains[39], they were first readily derived from 129S1/SvImJ mice ("129S"), leading to the establishment of widely used cell lines such as CJ7[38]. Wild-derived strains originate from individuals captured from wild mouse populations that were then inbred to homozygosity. WSB/J (*Mus musculus domesticus*, "WSB"), PWK/PhJ (*Mus musculus musculus*, "PWK"), and CAST/EiJ (*Mus musculus castaneus*, "CAST") represent the three main *Mus musculus* subspecies that diverged approximately one million years ago[56]. Each wild-derived strain has distinct genetic and phenotypic traits. For example, CAST mice are immune to flaviviruses but highly susceptible to other viruses such as orthopoxviruses and influenza A, while PWK show resistance to influenza A and sex-specific responses to diet-induced obesity. Despite typical fertility rates, WSB mice have significantly reduced sperm count and altered sperm morphology[54]. Together, these strains encompass approximately 23 million unique SNP sites and 350 million base pairs of structural variation[39,48]. The genetic diversity among the CC and DO founders exceeds that of the current humans[56].

Here, we characterize gene expression in the 8 founder strains of the Collaborative Cross and Diversity Outbred panels across 8 distinct tissue groups: (1) cortex and hippocampus, (2) diencephalon and pituitary gland, (3) muscle (gastrocnemius), (4) heart, (5) liver, (6) kidney, (7) adrenal gland, and (8) male and female gonads. We sample 4 adult males and females per genotype in each tissue and recover 108 heterogeneous cell types and states. We uncover strain-specific differences in cell type prevalences and cell states. Understanding the cell type-specific genetic regulation of gene expression is crucial for deciphering the mechanisms underlying complex traits. By leveraging single-nucleus RNA-seq, we are poised to unlock a deeper understanding of how genetic variation shapes cellular identity and function.

4.3 Results

Variation in body and tissue weight based on genotype and sex

We comprehensively collected 8 coordinated tissue groups across 8 diverse genotypes from young adult mice aged 10 weeks for a minimum of 4 males and females for each of the 8 founder strains. Wherever possible, the same set of 64 individuals are profiled for each tissue resulting in a coordinated whole-body dataset in each mouse. Our core dataset thus consists of 8 tissues x 8 strains x 8 reps/strains, for a total of 512 samples, which we call the "8-cube" dataset (Fig. 4.2a). We collected detailed metadata during sample collection including body and tissue weight, collection times, and estrus stage in females. As expected, NZO is the heaviest, weighing 1.95 times more than the average weight of 22 grams (g) across all strains (Fig. 4.1a-i). Whereas NZO is 43 g on average, the wild-derived strains are typically the smallest with a mean weight of 14.4 g. Sexual dimorphism is apparent in both body weight (with males weighing 16% more than females on average within each strain) as well as in certain tissues such as kidney and gastrocnemius muscle, where male tissues weigh 20.5% and 31.1% more than female tissues within each strain, respectively. Females and males are roughly the same size in AJ and NZO while B6 and NOD have the most sexual dimorphism in terms of body weight, with males weighing almost 1.5 times more than females. The correlation of body weight to tissue weight ranges from a Pearson R^2 to 0.94, with brain regions having the lowest (Fig. 4.1a, b) and kidney and liver having the highest correlation (Fig. 4.1h, i). As tissue size increases, the correlation between body weight and tissue weight strengthens, with most tissues exhibiting a positive correlation (> 0.4), except for brain regions and adrenal glands. In summary, sexual dimorphism based on body weight varies across genotypes and tissue weight generally scales with body weight depending on the tissue type.

We identify 108 unique cell types across 5.9 million nuclei

We extracted the nuclei from each tissue of our 8-cubed dataset, with a few repeats increasing the total number to 515 samples. As we have 64 samples per tissue, we performed combinatorial cell barcoding for two tissues in each split-pool experiment[88] to detect experimental batch effects and maximize sample multiplexing (Methods). In the first 8 out of 12 columns of our 96 well first barcoding plate, each sample from one tissue is loaded in its own well. In the remaining 4 columns, two samples from the second tissue of the same sex and replicate but distinct genotypes are multiplexed in a single well in 32 wells. The chosen genotype pairs to multiplex were based on calculating pairwise Hamming distances between strains based on 1,537,904 SNP regions followed by a maximal weight matching algorithm. This approach pairs the most distinct strains to each other, resulting in CAST multiplexed with 129, NOD with B6, PWK with AJ, and WSB with NZO (Fig. 4.2b). These wells need to be genetically demultiplexed to assign the mouse of origin correctly. This pattern was repeated for all 8 tissues (Fig. 4.2c). Each plate corresponds to an experiment with an expected yield of 1 million nominal nuclei conducted using Parse Biosciences Evercode Mega kits[89] (Methods). Thus, each tissue for each mouse replicate is sequenced twice – once as a single well and a second time mixed with one other genotype (Fig. 4.2c). We loaded approximately 1 million nuclei per tissue, which we sequenced with 20 billion reads for a depth of 20,000 reads per nucleus. The complete dataset consists of 8 million nominal nuclei sequenced with 160 billion short reads.

After quality control filtering (Methods), we recovered 5.9 million nuclei, 640,918 of which are from B6: 713,239 nuclei from muscle, 725,168 from liver, 617,453 from kidneys, 488,092 from ovary and oviduct, 256,090 from testes and epididymis, 612,840 from cortex and hippocampus, 814,531 from diencephalon and pituitary gland, 539,011 from adrenal glands, and 497,703 from heart. A principal component analysis of the tissues by grouping all nuclei from each individual into a pseudobulk gene expression matrix reveals high replicate concordance and expected clustering patterns between tissues (Fig. 4.2d). PC1 (33.23% of the variance) separates brain regions from other tissues, PC2 (20.02% of the variance) separates

liver and kidney from other trunk tissues, and both PC2 and PC3 (12.75% of the variance) separate male gonads from other tissues. These results agree with previous studies that have also shown a distinct transcriptional signature in liver and brain that distinguish them from other trunk organs[84], and that testis expresses the most tissue-specific genes when compared to 17 other tissues[274].

We annotated each tissue separately for a combined total of 363 clusters and 108 subtypes and states, compared to the 53 subtypes and states detected in the ENCODE4 (B6CAST F1) mouse dataset at the same 2-month timepoint in 4 of the same tissues. We recovered comparable proportions of major cell types across previously characterized tissues. Most tissues contain a primary cell type such as neurons in the brain, myonuclei in skeletal muscle, and cortical cells in adrenal gland (Fig. 4.2e). As in ENCODE4, the heart is roughly split by thirds into cardiomyocytes, fibroblasts, and endothelial cells. Most newly added tissues are also dominated by a major cell type: hepatocytes in liver, spermatocytes in male gonads, and epithelial cells in kidney (Fig. 4.2e). Evaluation of the number of genes detected across general cell types reflects similar trends as in ENCODE4, where neurons have the largest nuclear transcriptomes (Fig. 4.2f). Increased sequencing depth also increased the median number of genes detected in neurons and adipocytes, while genes detected in other cell types such as myonuclei, adrenal cortex, and endothelial cells remain similar to ENCODE4. Variations in cell type ratios and gene detection compared to ENCODE4 could stem from technical factors or illustrate how genetic background shapes cellular characteristics, as the ENCODE4 data is from an F1 cross. Further investigations using F1 mice (crosses between B6 females and each of the other seven founder strains) will help elucidate the underlying reasons for these differences.

Enhanced granularity elucidates minor cell types

Of the minor cell types (<25% of the tissue), 12 are shared between two or more tissues including immune cells and endothelial cells (Fig. 4.2g). Other shared cell types include

adipocytes in all tissues except brain and liver and Schwann cells shared between heart and muscle. The increased number of captured nuclei resolved three subtypes that did not form their own clusters in ENCODE4: myelinating and non-myelinating Schwann cells as well as lymphatic endothelial cells in skeletal muscle (Fig. 4.11a). Myelinating Schwann cells form a myelin sheath around axons of motor neurons to facilitate transmission of nerve impulses that control muscle contraction[275]. Non-myelinating Schwann cells, also called perisynaptic Schwann cells, cap the motor nerve endings at the neuromuscular junction (NMJ)[275]. They have been shown to maintain synapse stability and regulate synapse plasticity, aiding in repair after injury or weakening[275]. While myelinating Schwann cells are marked by upregulation of *Mpz*, *Pmp22*, and *Prx*, non-myelinating markers include *Scn7a* and *Slc35f1*[276] (Fig. 4.11b). We recover 2,447 myelinating Schwann cell nuclei, or 0.34% of the entire muscle dataset, and 1,175 non-myelinating Schwann, or 0.16% (Fig. 4.11a). Thus our sequencing of a greater number of nuclei achieved the granularity needed to recover these minor yet essential cell types.

Brain regions have the most cell type diversity across all tissues

Despite having the highest number of distinct cell types compared to all other tissues, brain regions display the least genotype-specific clustering. The cortex (CX) and hippocampus (HC), which we sequenced separately in ENCODE4, are involved in learning and memory, which are critical functions for the survival of the organism in the environment[277]. The CX/HC clusters exhibit the most even distribution across genotypes and sexes of all our tissues (Fig. 4.3a). We identified 25 cell types distributed across 41 clusters. CX/HC cell types also exhibit the least proportional variation by genotype compared to all other tissues (Fig. 4.3a, b, c, d). However, we noted some genotype-driven cluster with higher clustering resolutions in mature oligodendrocytes of CX/HC, as well as a cluster composed of PWK and CAST nuclei alone in diencephalon/pituitary (Fig. 4.4a). Oligodendrocytes are responsible for myelination, the process of forming insulating myelin sheaths around

axons of neurons to increase the speed and efficiency of signal conduction[196]. Regulatory topic modeling using Topyfic[128] in 61,368 CX/HC oligodendrocytes identified one PWK and CAST-specific topic out of the 9 recovered topics (Fig. 4.3e, f) (Methods). Some of the highly weighted genes specific to topic 3 are known to play crucial roles in mammalian neurogenesis, including *Cic*, *Ski*, and *Med12*[278–280]. Additionally, oligodendrocyte-specific expression of top-weighted genes like *Foxo3* and *Ptma* have been associated with multiple sclerosis in humans[281,282]. While clustering all nuclei from a given tissue reveals robust signatures, particularly if genotype and/or sex influences gene expression in the primary cell type, the results in oligodendrocytes (15.8% of the total number of nuclei in the CX/HC dataset) suggest that conducting analyses within individual cell types may uncover additional genotype-associated transcriptional variation.

The diencephalon is located between the cerebral hemispheres and the brainstem and includes the thalamus, hypothalamus, epithalamus, and subthalamus[283]. Though much smaller than the thalamus, the hypothalamus regulates essential bodily functions such as hunger and stress responses through specialized clusters of neurons (also called nuclei), such as the corticotropin-releasing hormone (Crh) producing neurons in the paraventricular nucleus[284]. The hypothalamus communicates with the pituitary gland through the release of hormones such as Crh to form systems such as the hypothalamic-pituitary-adrenal (HPA) axis and the hypothalamic-pituitary-gonadal (HPG) axis that control stress and reproductive functions, respectively[285,286]. While we focused our dissection on recovering primarily the hypothalamus, we also included neighboring regions of the thalamus, which are characterized by high *Tcf7l2* expression, as well as the pituitary gland. Of the 23 cell types recovered in diencephalon/pituitary, 6 are neuroendocrine cell types found in the pituitary gland (Fig. 4.4a, b). Even smaller than hypothalamus, the pituitary gland is anatomically divided into three main sections[287]. The anterior lobe constitutes the bulk of the gland and contains five primary hormone-secreting cell types: somatotropes, lactotropes, thyrotropes, corticotropes, and gonadotropes[287]. These cells produce growth hormone (Gh), prolactin

(Prl), thyrotropin (Tsh), adrenocorticotropic hormone (Acth), and gonadotropins (including Fsh), respectively[287] (Fig. 4.4b). Out of all other celltypes in diencephalon/pituitary, only mature oligodendrocytes show evidence of genotype-driven clustering, but are represented evenly across the genotypes (Fig. 4.4c, d). Among other tropes, melanotropes have apparent differences in proportion by genotype, particularly enriched in 129. The 129 diencephalon/pituitary dataset contains 2.9% melanotropes compared to 1.3% on average in the rest of the genotypes. Melanotropes are found in the intermediate lobe of the pituitary, or pars intermedia[288]. They generate melanocyte-stimulating hormone (Msh) from pro-opiomelanocortin, a precursor protein that undergoes post-translational processing to produce various peptides, including Acth and beta-endorphin[288]. Once released into the bloodstream, Msh can travel to other tissues where it exerts effects including stimulating melanocytes to produce melanin, the pigment responsible for coloration in skin, hair, and eyes[288,289]. Msh also has immunomodulatory effects and plays a role in the HPA axis, interacting with Crh and Acth to regulate the secretion of stress hormones[288]. Regulatory topics modeling in 12,204 melanotropes of the diancephalon/pituitary revealed 8 topics, most of which display enrichment in particular genotypes and/or sexes (Fig. 4.4e,f). Notably, B6 is strongly enriched in topic 3, while 129 is split between topic 5 and female-specific topic 7. Some of the genes shared in topic 5 and 7 include *Etv5*, *Lmna*, and *Xbp1*, while genes specific to topic 7 include *Xist*, *Mir224*, and *Npm1*. Sex-specific expression in *Pomc*-expressing cells has been previously observed, with males co-expressing *Pomc* and *Gh*, and females co-expressing *Pomc* and *Prl*[290], a pattern also shown in our data. However, prior single-cell studies have not concentrated on exploring sex differences in melanotropes specifically[290,291]. Notably, these studies were conducted in CD1[290] or B6[291], whereas we observe sex-specific topic enrichment in 129, AJ, and NOD.

Genotype-specific clustering in multiple cardiac cell types

Cardiomyocytes (27.0%, atrial and ventricular), endothelial cells (19.9%), and fibroblasts

(26.6%) constitute the bulk of heart tissue. Both ventricular cardiomyocytes and fibroblasts show evidence of genotype-specific clustering driven by PWK (Fig. 4.5a). Unique to cardiac fibroblasts out of all other celltypes except hepatocytes, B6 shows enrichment in a particular cluster (cluster 31, 34,395 nuclei). PWK and CAST share ventricular cardiomyocyte clusters and are almost completely absent in *Mus musculus domesticus* clusters, suggesting a strong association with subspecies. Despite the relatively equivalent size of the endothelial cell population to cardiomyocytes and fibroblasts, all four endothelial clusters are evenly distributed across the genotypes. This observation is shared in other tissues such as skeletal muscle, kidney, and male gonads.

Due to the high proportion of immune cells in the heart, subtypes such as B and T cells cluster separately from macrophages and are identified by marker gene expression (Fig. 4.5b). The largest shift in genotype distribution is in B cells, where PWK and CAST have the highest proportion compared to other genotypes and relative to their total number of nuclei (Fig. 4.5c). Proportions of major cell types are consistent across all strains (Fig. 4.5d). In order to further elucidate genotype-specific signatures, we performed regulatory topics modeling using Topyfic in 116,325 ventricular cardiomyocytes and recovered 10 topics (Fig. 4.5e, f). Distinct topics are enriched for PWK and CAST. The two topics enriched in CAST are also sex-specific; topic 6 is used more in males compared to topic 10 in females. Highly weighted genes in topic 6 include *Bhlhe40* and *Dbp*, known circadian rhythm factors[292,293], while highly weighted genes in topic 7 (enriched in PWK) include *Srebf1*, *Gtf2e1*, and *Gtf2e2*. Interestingly, CAST mice have been shown to demonstrate an "early runner" phenotype based on the timing of wheel-running relative to light/dark cycles[294]. Prediction of circadian rhythm based on expression of known circadian genes could help elucidate genotype-, sex-, and/or celltype-specific patterns in our data.

Differential proportions of sex-specific layers in the adrenal zona fasciculata across genotypes

In stark contrast to other celltypes, particularly in the brain, major celltypes in adrenal glands display significant sex-specific gene expression (Fig. 4.6a). Adrenal glands, paired endocrine organs resting on top of the kidneys, are responsible for hormone production[295]. The adrenal cortex synthesizes hormones such as glucocorticoids, aldosterone, and some sex hormones to regulate metabolism, electrolyte balance, and reproductive functions[295]. The adrenal cortex is subdivided into the zona glomerulosa (ZG) layer that secretes aldosterone (*Cyp11b2*) and zona fasciculata (ZF) that produces corticosterone[295] (Fig. 4.6b). The adrenal gland plays a pivotal role in orchestrating systemic responses through the hypothalamic-pituitary-adrenal axis[285]. The HPA axis orchestrates a hormonal cascade, beginning with signaling from the hypothalamus which releases Crh[285]. Crh stimulates the pituitary gland to secrete the adrenocorticotropic hormones, which in turn signals the adrenal cortex to produce glucocorticoids, including corticosterone[285]. Finally, the inner adrenal medulla secretes epinephrine (adrenaline) and norepinephrine in response to stress with widespread effects on the body such as heart rate and blood pressure[296]. Integrative analyses between diencephalon/pituitary and adrenal glands hold promise to unveil deeper insights into the comprehensive regulation orchestrated by the HPA axis.

We recovered the same cell types as in the ENCODE4 adrenal gland dataset in similar proportions (14 total). The majority of the nuclei belong to the ZF with a smaller amount associated with the ZG (Fig. 4.6a). We can distinguish epinephrine-producing chromaffin cells of the adrenal medulla from norepinephrine-producing cells based on expression of *Pnmt*, the primary enzyme that converts norepinephrine to epinephrine[296]. Previous studies identified the X-zone, a female-specific cortical layer that arises during puberty with marker genes *Akr1c18*, *Pik3c2g*, and *Thrb*[107]. Characterization of gene expression and transcription factor activity in adrenal gland in ENCODE4 also revealed male-specific patterns in the zona fasciculata that we annotated as "male-only ZF". While consistent proportions of male-specific ZF are observed across the 8 genotypes, remarkable variations are detected in the X-zone composition (Fig. 4.6b). Very few nuclei from B6 and NOD females fall into the

clear X-zone cluster identified by high expression of canonical marker genes (Fig. 4.6b, c). In contrast, CAST and NZO display substantial enrichment in the X-zone cluster relative to the total number of nuclei recovered from the genotypes (Fig. 4.6c, d). Comparing the ENCODE4 dataset, the proportion of the X-zone at the 2-month timepoint in B6/CAST F1 females (3.9% of all 2 month nuclei) is roughly halfway between the proportion in CAST shown here (7.0%) and several fold higher than B6 (0.6%). This observation may reflect that hybrids will sometimes exhibit an average phenotype between the two parents. However, these results are also surprising given that several experiments tracking X-zone dynamics are conducted in B6[107,258,259]. Planned IGVF experiments conducted in the F1 hybrids will help elucidate the technical or biological origin of this finding.

Topic modeling in 26,638 X-zone nuclei (Fig. 4.6 e, f) reveals topics highly specific to PWK and CAST in addition to NZO and WSB. Some of the top weighted and highly specific genes in NZO topic 1 compared to other genotype-specific topics 3, 5, 6, 7, and 9 include *Esr1*, *Rarb*, and *Bcl3*. In PWK topic 3, top genes include *Nfatc2*, *Aff3*, and *Padi2*. Two topics were enriched in both WSB and CAST with top genes *Dach1*, *Ppm1d* and *Pak1*, and *Irf8*, *Bcorl1*, and *Maf*, in the topics specific to each genotype, respectively. Of these genes, *Dach1* and *Padi2* have interesting implications in the adrenal gland. *Dach1* has been shown to inhibit aldosterone secretion in human adrenals and serves as a zona glomerulosa marker[297]. While *Padi2* has no direct evidence in adrenal gland, ablation of this gene in mice caused a delayed onset of puberty and had consistently lower serum testosterone levels[298]. Although mouse adrenals do not synthesize *Cyp17a1* and thus do not secrete androgens such as testosterone[202], from our ENCODE4 postnatal mouse study we found X-zone dynamics are intricately linked to puberty.

Sexually dimorphic gene expression in cell types of kidneys and liver

Located directly below the adrenal glands, kidneys serve as primary excretory organs that filter blood to maintain fluid and electrolyte balance in the body[299]. The basic filtration unit

in the kidney is the nephron, a complex structure consisting of several distinct segments[299]. The nephron comprises the renal corpuscle, which includes Bowman's capsule, and a tubular system consisting of the proximal convoluted tubule, loop of Henle, distal convoluted tubule, and collecting duct[299,300]. Blood is initially filtered through the glomerulus, a network of capillaries within Bowman's capsule[301]. The glomerular basement membrane is made up of podocytes, endothelial cells, and basement membrane, where the foot-like podocyte projections form a size-selective barrier[301]. The filtrate then passes through the epithelial cell-lined proximal convoluted tubule where reabsorption of water and ions occurs[299]. Subsequently, the filtrate enters the loop of Henle, which establishes the osmotic gradient within the kidney to facilitate water reabsorption[299,302]. The loop of Henle, also lined by specialized epithelial cells, consists of a descending limb where water is reabsorbed and an ascending limb where ions are actively transported[299,300]. Following the loop of Henle, the filtrate enters the distal convoluted tubule where further reabsorption and secretion of ions occur[299]. The filtrate passes into the collecting duct, where additional water and ion reabsorption is carried out by principal cells, while intercalated cells are responsible for acid-base balance[299,300]. The filtrate is then transported to the bladder for excretion.

We detect epithelial cells of all segments of the nephron as well as specialized cells such as podocytes. The largest clusters correspond to epithelial cells of the proximal (42.9% of the dataset) and distal (4.9%) tubules, loop of Henle (16%), and principal cells (4.9%) and intercalated (4.1%) cells of the collecting duct (Fig. 4.7a). Endothelial cells and fibroblasts also make up a substantial fraction of the total dataset at 10.3% and 4.9%, respectively. One small cluster making up less than 1% of the total kidney nuclei can only be identified by expression of common epithelial markers and additional marker *Megf11* (Fig. 4.7b). A previous single-nucleus study in mouse kidney[303] also identified this *Megf11*+ cluster as well as an *Ncam1*+ cluster, which in our data also expresses *Cp* and *Wt1* and thus most likely consists of parietal epithelial cells (PSCs), an important glomerular cell type[304]. Additional *Megf11*+ cluster markers include *Erbb4*, *Pax2*, and *Prox1*, which may indicate a subtype

of loop of Henle epithelial cells[305,306]. All kidney cell types apart from fat cells, likely a dissection artifact, are evenly distributed across genotypes, although PWK and CAST have a smaller proportion of proximal tubule epithelial cells than others (Fig. 4.7c, d).

Previous single-cell transcriptome studies in the kidney identified sexual dimorphism in the proximal tubule[300], a finding supported by our data. The majority of our proximal tubule nuclei (71%) fall into sex-specific clusters, defined as those comprising over 90% of one sex, with approximately a third (27.8%) showing genotype-specific clustering, characterized by over 90% representation of a single genotype (Fig. 4.7a). The majority of genes previously shown to be differential by sex in proximal tubules[300] are also identified in our data (Methods). For example, the prolactin receptor (*Prlr*), nephronectin (*Npnt*), prominin-1 (*Prom1*), solute carriers (*Slc22a29*, *Slc39a5*, and *Slc6a18*), and of course *Xist* are upregulated in females of all genotypes. In males, alcohol dehydrogenase (*Adh1*), cytochrome P450 enzymes (*Cyp2e1*, *Cyp2j13*, *Cyp4a12a*, and *Cyp4b1*), solute carriers (*Slc22a28*, *Slc22a30*, *Slc7a13*) and steroid dehydrogenases (*Hsd17b11* and *Hsd3b3*) are upregulated in all genotypes. Thus, we decided to apply regulatory topics modeling to 264,996 proximal tubule epithelial cells (Fig. 4.7e, f). Out of 11 topics, 5 are sex-specific and one, topic 2, is enriched in PWK and CAST (Fig. 4.7e). Topic 2 is not enriched in either sex, indicating a non-*Mus musculus domesticus* cellular program independent from sex. Some of the top-weighted regulatory genes in our genotype-specific proximal tubule topic include genes that are mutated (*Brca2*, *Fancm*) or serve as biomarkers (*Riox2*) in kidney cancer[307–309]. In summary, our kidney dataset supports the sex-driven transcriptional differences noted previously in proximal tubules in addition to differences depending on genotype that are both linked to and independent from sex.

The liver plays a crucial role in maintaining homeostasis by regulating various metabolic processes[310]. The functional units of the liver are lobules[310,311]. Each lobule is a hexagonal arrangement of hepatocytes radiating outward from a central vein[310]. Blood from the hep-

atic portal vein, rich in nutrients and toxins absorbed from the digestive tract, enters the lobule through sinusoids, specialized capillaries lined with endothelial cells, Kupffer cells, and stellate cells[310,312]. As blood traverses through the sinusoids, hepatocytes facilitate metabolism of nutrients, detoxification, and secretion of bile[310]. The network of bile ducts, lined with cholangiocytes, collects and transports bile synthesized by hepatocytes towards the bile ductules and ultimately to the gallbladder for storage or to the intestines for aiding digestion[310,313]. We detect 14 cell types in liver, the majority of which (79% of the total nuclei) are hepatocytes (Fig. 4.8a, b). Of the 574,931 hepatocyte nuclei, 100% and 70% are in sex-specific and/or genotype-specific clusters, respectively (over 90% consisting of one genotype or sex) (Fig. 4.8a). While sexes are for the most part evenly distributed across all non-hepatocyte clusters, we note enrichment of female B6 and AJ in a cluster of cycling cells (Fig. 4.8a, c). AJ and NZO have a significantly higher proportion of Kupffer cells than other genotypes (Fig. 4.8a, c). As the resident macrophages of the liver, Kupffer cells play a crucial role in eliminating pathogens such as bacteria that enter the bloodstream through the gastrointestinal tract[314]. In NZO, 5.2% of nuclei are from Kupffer cells, double the average in all other genotypes (2.1%) (Fig. 4.8d). Metabolic stresses that promote insulin resistance and type 2 diabetes (T2D) also activate inflammation- and stress-induced signaling pathways, resulting in chronic inflammation in tissues including the liver[315]. Given the propensity of NZO to develop T2D[33] and its early-onset obesity at 2 months (Fig. 4.1), we hypothesize that this increase in the proportion of Kupffer cells may be intricately linked to early progression of T2D.

Applying regulatory topics modeling to hepatocytes results in 11 topics, 3 of which show strong sex specificity (Fig. 4.8e, f). As in kidney, the genotype-specific topics are independent from sex. Despite genotype-driven clustering in the full dataset (Fig. 4.8a), only 2 topics are clearly genotype-specific (PWK topic 6 and CAST topic 11) (Fig. 4.8e). Comparing female topic 2 to the two male topics 3 and 8 reveals several known female-specific transcriptional regulators, such as *Cux2*, *Tox*, and *Trim24*[316]. In the liver, *Cux2* acts as a

master regulator by activation of female-specific target genes including *Tox* and *Trim24*[316]. Genes with significant weights in male-specific topics include *Bcl6*, *Stat5b*, *Ppargc1b*, *Smad3*, *Nr1d1*, and *Nr1d2*. These genes have previously been linked to male-specific patterns of gene and/or protein expression in the liver[317–320]. *Bcl6* controls masculinization of hepatic gene expression during puberty, enhancing survival in male mice during severe bacterial infection[317]. However, it also contributes to fatty liver and glucose intolerance under conditions of dietary excess, leading to a male predisposition for these conditions[317]. In PWK topic 6, top-weighted genes *Mir22* and *E2f7* have both been found to act as tumor suppressors in liver cancer[321,322]. Conversely, *Atf7* appears exclusively in CAST topic 11, and has been implicated in epigenetic regulation of gene expression in mouse liver in response to metabolic changes induced by diet[323]. To summarize, the analysis of hepatocyte topics uncovers both established sex-specific patterns and novel genotype-specific expression signatures.

The testes/epididymis dataset captures dynamic stages of spermatogenesis

We identify 14 distinct cell types in testes and epididymis, with spermatogenic cells comprising 39% of the total recovered nuclei (Fig. 4.9a, b). Unlike other tissues where a differentiated primary cell type also makes up the majority of the tissue, here we can detect the dynamic stages of its differentiation process. Spermatogenesis is the continuous process of sperm production in males that begins during puberty and continues throughout adulthood[324]. It starts with spermatogonia, which are the stem cells in the testes[324]. These cells undergo mitosis to produce primary spermatocytes, which then undergo meiosis to form secondary spermatocytes[324]. Further division results in spermatids, which undergo a process of maturation called spermiogenesis, during which they develop into spermatozoa, or sperm cells[324]. We capture the most cells at the spermatid phase, followed by mature sperm, spermatocytes, and spermatogonium (59%, 18%, 16%, and 7% of the total germ cell population) (Fig. 4.9a). In the epididymis, principal cells play a crucial role in the maturation and storage of sperm[325]. These cells are specialized according to the segment of the

epididymis they inhabit, with distinct gene expression in each region[325] (Fig. 4.9b). We recover 10,929 principal cells of the caput, 22,006 of the corpus, and 18,667 of the cauda (21%, 43%, and 36% of the principal cell population) (Fig. 4.9b, c). Overall, major germ cell and principal cell populations are represented evenly across genotypes (Fig. 4.9c, d). Of the minor cell types, Sertoli cells appear enriched in WSB (22% vs. 14% average in other genotypes) while Leydig take up a larger proportion in NZO (10% vs. 5.5% average in other genotypes) (Fig. 4.9d). Located within the seminiferous tubules of the testes, Sertoli cells provide structural support to developing germ cells and are regulated by follicle-stimulating hormone from the pituitary gland as part of the HPG axis[324]. In contrast, Leydig cells found in the tubule interstitium produce testosterone in response to luteinizing hormone from the pituitary, thereby regulating male secondary sexual characteristics and spermatogenesis[324]. The notable increase in Leydig cells in NZO is intriguing considering the association between obesity in males and decreased testosterone levels[326]. This heightened presence of Leydig cells might serve as a compensatory mechanism to counteract the impact of obesity on testosterone production. Due to the abundance of spermatogenic cells in the dataset and their key role in reproduction, we decided to apply topics modeling to 18,111 mature sperm cells. We recovered 12 regulatory topics, two of which are specific to CAST (topic 6) and PWK (topic 8), while topic 1 and to a lesser extent topic 12 are enriched in both non-*Mus musculus domesticus* strains (Fig. 4.9e, f). Some of the top-weighted genes shared in CAST topic 6 and non-domesticus topic 12 include *Bcl9l*, a cofactor in Wnt/Beta-catenin signaling[327], and *Sin3b*, overexpression of which has been shown to promote the formation of microcephalic sperm in a human study[328]. *Kdm4b*, a histone demethylase associated with spermatogenesis, is highly weighted in non-domesticus topic 1, 12, and PWK topic 8[329]. In summary, the male gonads dataset facilitates the examination of the dynamic stages of spermatogenesis, where genotype may impact gene expression at each developmental stage, as demonstrated here in mature spermatozoa.

***Mus musculus domesticus* variation in ovarian theca**

We collected both ovaries and oviducts along with estrus stage (Methods) and detect 14 cell types in the female gonads dataset (Fig. 4.10a, b). The ovaries are the primary cite of oocyte production and hormone secretion and consist of specialized cell types such as theca and granulosa cells[330]. Theca cells are in the outer layer of the ovarian follicles and are involved in producing androgens, which are precursors to estrogen[330]. Granulosa cells are found in the inner layer of the follicles, producing estrogen and supporting oocyte development and maturation[331]. The oviducts, also known as fallopian tubes in humans, are lined by ciliated and secretory endothelial cells that facilitate the transport of oocytes[332]. Unlike other tissues, major cell types make up even proportions of the total number of nuclei, with 25% granulosa, 18% theca, 14% secretory and 8% ciliated endothelial. The ovarian stroma fills in 10%. The remaining 15% of nuclei are made up of vascular endothelial, epithelial, smooth muscle, and immune cells. Although the granulosa cells are the predominant cell type in ovary, theca cells display more genotype-driven clustering, with 27% of theca nuclei falling into a genotype-specific cluster (Fig. 4.10a). Thus, we performed topics modeling in 86,282 ovarian theca cells and recovered 10 regulatory topics. Interestingly, we detect strong topic enrichment in *Mus musculus domesticus* strains WSB and NZO. The top-weighted gene in NZO topic 9, *Mir218-1*, is embedded in the *Slit2* host gene. Both the microRNA and its host have been implicated in ovary, with *Mir218* associated with anti-cancer effects and Slit2, a ligand in the Slit-Robo signaling pathway, associated with degradation of the corpus luteum[333,334]. We detect a cluster of nuclei that appear to be derived from the corpus luteum, characterized by specific expression of the prostaglandin F receptor (*Ptgfr*)[335] and steroidogenic acute regulatory protein (*Star*)[336] specifically enriched in B6 and AJ with very little contribution from other genotypes (Fig. 4.10a, b, c, d). These proportional differences may be associated with the estrus stage, which is closely linked to genotype in our dataset. This association arises from the shared estrus cycle observed among females of the same genotype during sample collection, possibly a result of being housed together. The corpus luteum forms from the collapsed pre-ovulatory follicle after ovulation, which occurs

just after the end of estrus[337]. Metestrus follows estrus, when the corpus luteum begins to develop[337]. It eventually declines during diestrus[337]. Therefore, we may be detecting the early corpus luteum as a distinct cluster. In conclusion, we observe robust genotype-specific expression patterns that may also be influenced by estrus stage, which must be considered when evaluating differences in gene expression and cell type proportions.

Impact of subspecies on gene expression in type 2 myonuclei

As is observed in multiple tissues, a predominant cell type, specifically myonuclei, dominates the majority of clusters in skeletal muscle. Subspecies significantly influences clustering in type II myonuclei, with PWK and CAST often clustering separately from *Mus musculus domesticus* strains (Fig. 4.11a). Some clusters are specific to WSB in major cell types, such as type IIb skeletal muscle cells. The gastrocnemius, or calf muscle, primarily consists of fast-twitch (type II) fibers with a smaller population of slow-twitch (type I) fibers[221]. Subtypes of type IIa fibers are distinguished by expression of myosin heavy chain protein isoforms, with type IIa corresponding to high expression of *Myh2*, type IIb to *Myh4*, and type IIx to *Myh1*[112,113] (Fig. 4.11b). Additionally, type I and IIa oxidative fibers primarily use aerobic respiration for energy production, while type IIx and IIb glycolytic fibers rely on anaerobic glycolysis[338]. Glycolytic fibers fatigue faster than oxidative fibers due to the lower ATP production per cycle in anaerobic glycolysis[338]. Although type IIb fibers are evenly represented across the genotypes, CAST have a much smaller proportion of type IIa, while type IIx is enriched in both CAST and PWK (Fig. 4.11c, d). Interestingly, CAST exhibits resistance to disuse-induced muscle atrophy compared to other strains, particularly NOD and NZO[339]. CAST mice with an immobilized limb remained active and lost the least body weight during the experiment, especially compared to AJ, which experienced the greatest weight loss and decreased activity[339]. CAST also displayed potential fast-to-slow fiber type switching in the gastrocnemius upon unloading, indicated by increased expression of type I marker *Myh7*[339]. The lack of type IIa in CAST suggests that inherently low levels

of oxidative fast-twitch fibers may contribute to this strain-specific fast-to-slow phenotype. The transition to slow fibers during unloading presumably increases oxidative metabolism overall, as slow fibers are oxidative. This change in fiber type proportion may counteract the fatigue-prone type IIb glycolytic fibers, potentially aiding in maintaining the activity levels observed in CAST.

We performed topics modeling to investigate the regulatory programs driving the transcriptional variation in type II myonuclei. We recovered 9 subtype- and/or genotype-specific topics in 523,897 nuclei (Fig. 4.11e, f). Several topics also display sex specificity. Topic 7 shows enrichment in CAST and PWK while topic 5, which is also type IIb-specific, is enriched in NOD (Fig. 4.11e, f). Some of the top-weighted genes in topic 7 compared to all other topics include *Klf7*, *Egfr*, *Asah1*, and *Lrif1*. *Egfr* inhibition has been associated with promoting an oxidative slow-twitch phenotype in mouse tissue and C2C12 cell line[340]. Although topic 7 is not specific to any particular subtype, it is slightly less enriched in type IIa compared to type IIb and IIx. Mutations in both *Asah1* and *Lrif1* have been associated with human diseases such as spinal muscular atrophy and facioscapulohumeral muscular dystrophy, although neither has evidence in mice[341,342]. Further investigation by directly comparing topics and cross-referencing known protein-coding variants in these strains with genotype-specific topic genes may help elucidate their role in type II myonuclei.

Satellite cell activation in AJ skeletal muscle

Thus far, we focused on major cell types or subtypes thereof in each tissue for deeper exploration through differential expression analysis and topics modeling. Some cell types exhibit surprising genotype-driven expression differences, as observed in oligodendrocytes. Hence, we opted to investigate a small yet significant cell type in skeletal muscle: satellite cells, constituting only 0.7% of the dataset. As the resident stem cells in skeletal muscle, satellite cells play a crucial role in muscle development and repair. Upon myofiber damage, satellite cells become activated and initiate transcriptional programs by expressing key myogenic

regulatory factors (MRFs) such as *Myog*, *Myod1*, *Myf5*, and *Myf6*[111]. Through asymmetric division, a fraction of *Pax7*+ satellite cells remains quiescent to sustain the stem cell pool[111]. Activated satellite cells persistently express myogenic genes and eventually fuse with damaged myofibers[111]. We recovered 7 topics, two of which (topics 3 and 4) showed enrichment in AJ mice while topics 5 and 6 were depleted in AJ and enriched in PWK, WSB, and 129S1 (Fig. 4.12a). AJ is known to be susceptible to muscular dystrophy due to dysferlin deficiency[37], caused by a 6,000 bp retrotransposon that spontaneously inserted itself into intron 4 of the gene, leading to splicing disruption[129]. Dysferlin is a protein of the sarcolemma associated with limb girdle muscular dystrophy 2B, Miyoshi myopathy, and distal anterior myopathy when mutated[129]. Previous studies in humans have shown that the proportion of activated satellite cells in dysferlinopathic muscle is higher than in control, but lower than in other myopathies[343]. The same study shows that dysferlin is upregulated in activated satellite cells of dystrophic muscle compared to control, despite the mutation causing an overall deficiency[343]. We detect upregulation of *Dysf* in AJ satellite cells, but downregulation in mature myonuclei compared to other genotypes (Fig. 4.12b). Differential expression analysis between genotypes in satellite cells reveals key regulatory genes specifically upregulated in AJ satellite cells, most notably *Myog* (log2 fold change of 2.7 and adjusted p-value of 0.0002) (Fig. 4.12c). Comparison of gene weights in our topics identifies *Pax7* as specific to topics 5 and 6 while *Mef2c* and *Mir133a-1* are highly weighted in AJ topics 3 and 4 (Fig. 4.12d). *Mef2c* synergizes with MRFs to activate myogenesis and *Mir133a* is a classic "myomiR", or microRNA whose expression is highly specific to skeletal muscle and required for muscle development[344,345]. Notably, some fraction of satellite cells in all genotypes participate in topics 3 and 4, although to a much lesser extent than AJ (Fig. 4.12e). This observation suggests that the activated program may not inherently be pathogenic, but rather excessively activated in AJ. A study conducted on satellite cells isolated from dystrophic mouse muscle demonstrated that their regenerative capacity remained intact compared to control muscle, implying that the *in vivo* environment plays a crucial role in regulating satellite

cell function[346]. Further exploration of other cell types that interact with satellite cells, such as fibro-adipogenic progenitors, may uncover additional genotype-driven changes in expression that influence the satellite cell microenvironment. In summary, our findings reveal over-activation of myogenic regulatory programs in AJ satellite cells compared to other genotypes, potentially reflecting its predisposition to early onset dysferlinopathy. While AJ is widely used as a mouse model across diverse research domains, including cancer and emphysema, researchers must remain aware of all phenotypes inherent to their selected genotype. Our findings demonstrate that even at 2 months, genotype-specific functional changes in gene expression occur in a critical cell type.

4.4 Discussion

Our 8-cube founder dataset of 5.9 million nuclei across 512 samples is a comprehensive map of transcriptional variation across diverse genotypes and both sexes, shedding light on the intricate relationship between genomic variation and gene expression regulation. This resource is particularly significant in the context of mouse research, where genetic background is often overlooked despite its profound impact on experimental outcomes. By systematically characterizing the transcriptional landscape across multiple genotypes and sexes, we highlight the importance of considering these factors in experimental design and interpretation. Notably, our findings challenge the common assumption that B6 mice are the prototypical mouse, as we detect substantial transcriptional variation even among commonly used laboratory strains. This underscores the necessity of considering genetic diversity in mouse studies, as it may significantly influence phenotypic traits and confound experimental results. Moreover, our study underscores the significance of sex as a biological variable in gene expression regulation. While sex differences in certain tissues have been well-documented, our analysis across diverse genotypes reveals sex-specific transcriptional patterns that extend beyond pre-

viously characterized tissues. By providing a coordinated dataset encompassing both males and females across various strains and subspecies, we offer a valuable resource for understanding the impact of sex on gene expression in diverse genetic contexts. Importantly, our findings highlight genotype-specific functional changes in critical cell types, even in healthy-looking, young adult mice where physiological symptoms are not yet apparent. For instance, we observed enrichment of Kupffer cells in NZO livers and over-activation of satellite cells in AJ muscle, indicative of transcriptional signatures associated with diseases like type 2 diabetes and muscular dystrophy, respectively. Furthermore, our study demonstrates the power of single-cell RNA-seq in elucidating cell type-specific transcriptional regulation with enhanced resolution. By profiling millions of nuclei across diverse tissues, we not only capture cell type interactions within individual tissues but also have the potential to uncover coordinated responses across different tissue systems, such as the HPA and HPG axes. Additionally, the large number of cells profiled enables the detection of minor cell types and cell states that may be overlooked or averaged out in bulk RNA-seq approaches. For example, we characterized satellite cells across the 8 strains, a cell type that comprises less than 1% of the muscle dataset, highlighting the enhanced granularity afforded by single-cell analysis. Overall, our dataset provides a foundational resource for advancing the systematic understanding of the genomic basis for cell type-specific transcriptional regulation.

While our study offers comprehensive insights into transcriptional variation across genotypes and sexes, several considerations highlight the need for transparency and interpretability. One notable challenge stems from the inadvertent association between estrus stage and genotype. For instance, we observed that PWK females are predominately in estrus, whereas CAST are in diestrus or proestrus. This inherent link between genotype and estrus stage complicates the differentiation of genotype-specific effects from those attributable to estrus stage, particularly in tissues where estrus stage is pivotal, such as the ovary. Although literature does not indicate estrus cycle synchronization among co-housed mice, we speculate that cohabitation of females from each genotype with males in neighboring cages may have

influenced the observed distribution of estrus stages. Additionally, the timing of tissue collection presents another challenge, especially concerning its impact on circadian rhythm. To streamline the collection process, females of each genotype were typically collected in the morning (9-11am), followed by males from 11am to 1pm. This sequential collection resulted in a circadian rhythm linked to sex, potentially confounding analyses of circadian-regulated genes. Moreover, the collection of NZO samples occurred later, with females starting at 11am and males ending at 4pm, introducing additional variability in circadian rhythms across genotypes. Additionally, our study's combinatorial barcoding approach and high degree of multiplexing result in somewhat less control over the number of nuclei recovered per sample compared to more traditional droplet-based methods. This variability in nuclei recovery is particularly evident in heart, where we recover nearly half as many nuclei in WSB (8.9%) than in NZO (16%) in the same dataset. Although efforts were made to address this variability through makeup experiments using leftover nuclei, it is crucial to exercise caution when comparing cell type proportions, always considering the overall number of nuclei recovered. While some are inherent to the experimental design, these complexities underscore the necessity of carefully addressing potential biases in downstream analyses.

Future integration of the founder snRNA-seq data presented here with an F1 "8-cube datasets (B6 females crossed with each of the seven founder strains plus a B6 control) will facilitate the development of allele-specific gene expression pipelines. This will enable precise mapping of *cis* and *trans* regulatory effects at the cell type level, elucidating the regulatory mechanisms governing gene expression. Expanding our snRNA-seq dataset to include 33 strains of the Collaborative Cross enables genome-wide mapping of expression quantitative trait loci (eQTLs) at a broad resolution. We aim to supplement single-nucleus chromatin accessibility data in selected or all tissues to refine eQTL loci. The insights gleaned from the founder dataset presented here lay a solid foundation for hypothesis generation in subsequent F1 and CC analyses. Leveraging this rich dataset and building upon our observations, we aim to deepen our understanding of the genetic architecture underlying complex phenotypic traits.

Ultimately, our findings will not only inform QTL mapping approaches in human data but also identify potential therapeutic targets in mouse models of human disease, thereby paving the way for targeted therapeutic interventions.

Acknowledgements

Thanks to Brian Williams for dissecting muscle tissues and conducting pap smears, Grant MacGregor for dissecting trunk tissues, and Shimako Kawauchi for dissecting brain tissues, determining estrus stage, and ordering mice from Jackson Labs. Thanks to Heidi Liang and Ghassan Filimban for assistance with nuclei isolation, barcoding, and library preparation. We also acknowledge Sina Booeshaghi for his maximal weight matching analysis between the genotypes and input on the experimental design, Delaney Sullivan for his development of the genotype demultiplexing package, and Lior Pachter for his guidance and insight throughout the data collection and analysis process. Finally, we thank the UCI Transgenic Mouse Facility for housing the mice and use of their facilities and UCI GRTH for sequencing the libraries.

Data and code availability

- Data availability: IGVF measurement sets are listed in Table S1.
- Data processing/figure generation code

4.5 Supplementary tables

- Table S1: IGVF portal measurement set metadata for all data used.
- Table S2: Mouse metadata.

4.6 Methods

Mice and tissue collection

Mice were ordered from Jackson Laboratories and housed at the UCI Transgenic Mouse Facility under controlled conditions. All animal procedures were approved by the Institutional Animal Care and Use Committee, protocol #AUP-21-106. Metadata for each animal and tissue, including mouse ID, sex, date of birth and euthanasia, time of euthanasia, dissector ID, body and tissue weights, and estrus stage are detailed in Table S1. Euthanasia was performed by anesthesia using isoflurane, followed by decapitation for the collection of whole blood in EDTA-coated tubes (BD cat. #367856). A pap smear was conducted on female mice before tissue collection and stored on Superfrost slides (Fisher Scientific cat. #12-550-15). Organs and tissues were dissected by three expert dissectors in parallel: brain regions (left and right cortex and hippocampus, cerebellum, and diancephalon and pituitary), trunk organs (heart, lungs, liver, adrenal gland, kidney, gonads, perigonadal fat, and brown adipose tissue), and specific hindlimb muscles (soleus, plantaris, gastrocnemius, tibialis anterior, and EDL). Trunk and muscle tissues were washed in ice-cold HBSS. Tissues were then flash-frozen in liquid nitrogen and biobanked at -80°C until further processing. Estrus stage was determined by crystal violet staining and observation under a light microscope.

Purification of nuclei from mouse tissues

Eight replicates of each tissue from each of the founder genotypes were processed per day. Flash-frozen tissues were transferred to a chilled gentleMACS C Tube (Miltenyi Biotec cat. #130-093-237) with Nuclei Extraction Buffer (Miltenyi Biotec cat. #130-128-024) supplemented with 0.2 U/ul RNase Inhibitor (NEB cat. #M0314L) on ice. Nuclei were dissociated from whole tissues using a gentleMACS Octo Dissociator(Miltenyi Biotec cat. #130-095-937). The suspensions were sequentially filtered through 70 um and 30 um strainers (Miltenyi Biotec cat. #130-110-916 and #130-098-458, respectively). Nuclei were then resuspended

in cold PBS (Life Technologies cat. #15260037) with 0.1% BSA (Life Technologies cat. #15260037) and 0.2 U/ul RNase inhibitor for manual counting using a hemocytometer and DAPI stain (Thermo cat. #R37606). Debris removal solution (Miltenyi Biotec cat. #130-109-398) was applied to gastrocnemius tissue, forming a density gradient to separate nuclei bands from debris layers. For most tissues, 4 million nuclei per sample were fixed using Parse Biosciences' Nuclei Fixation Kit v2 (Parse Biosciences cat. #) according to the manufacturer's protocol. For smaller tissues such as adrenal gland and female gonads, at least 1 million nuclei were used as input for fixation. Briefly, nuclei were incubated in fixation solution for 10 minutes on ice, followed by permeabilization for 3 minutes on ice. The reaction was quenched, then nuclei were centrifuged and resuspended in 300 uL Nuclei Buffer (Parse Biosciences cat. #) for a final count. DMSO was added before freezing fixed nuclei at -80°C.

Parse Split-seq experiments

Nuclei were barcoded using Parse Biosciences' WT Kit v2 (cat. #ECW02030), following the manufacturer's protocol. Fixed, frozen nuclei were thawed in a 37°C water bath and added to the Round 1 reverse transcription barcoding plate at 37,500 nuclei per well. The plate design alternated females and males across columns. A majority of the plate comprised a main tissue, where each individual sample was loaded into a unique well (64 wells in total for the 8 genotypes and 4 male and female replicates). The remaining 32 wells contained a multiplexed tissue, different from the main tissue, where two replicates were pooled from two distinct genotypes in the same well. In situ reverse transcription (RT) and annealing of barcode 1, then nuclei were pooled and distributed into 96 wells of the Round 2 ligation barcoding plate for in situ barcode 2 ligation. Finally, nuclei were pooled again and redistributed into 96 wells of the Round 3 ligation barcoding plate for in situ barcode 3 + UMI + Illumina adapter ligation. Finally, nuclei were counted using a hemocytometer and distributed into 16 subpools of 67,000 nuclei each. The nuclei in each subpool were lysed, and cDNA was purified using AMPure XP beads (Beckman Coulter cat. #A63881). The

barcoded cDNA then underwent template switching and amplification. After cleaning the cDNA using AMPure XP beads and performing quality checks with an Agilent Bioanalyzer, the libraries were prepared for Illumina sequencing. The cDNA samples were fragmented, size-selected using AMPure XP beads, and Illumina adapters were ligated. To create the final libraries, cDNA fragments underwent another round of amplification, adding the fourth barcode and P5/P7 adapters, followed by size selection and quality check with a Bioanalyzer (Agilent cat. #G2938A). Libraries were sequenced with two runs of the Illumina NovaSeq 6000 sequencer using the S4 Reagent Kit v1.5 300 cycle kit (cat. #20028312). All 16 subpools were sequenced together with 5% PhiX spike-in were sequenced to an average depth of total of 20 billion reads per experiment (approximately 20,000 reads per cell).

Read mapping and quantification

Cell-by-gene counts matrices were generated from NovaSeq fastqs using a custom in-house pipeline based on the kallisto bustools suite[95], `https://github.com/mortazavilab/parse_pipeline`. We used kallisto bustools to pseudoalign reads to the mm39 genome with GENCODE vM32 annotations, assign reads to genes, demultiplex cell barcodes, deduplicate UMIs[95]. The counts matrices, gene information, and cell barcodes were compiled into anndata H5ad files[99]. Auxiliary code merged detailed sample- and mouse-level metadata into the corresponding observation or "obs" table. Where appropriate, such as in this experimental design, the pipeline uses a custom genetic demultiplexing package klue (`https://github.com/Yenaled/klue`) to quantify the number of genotype-specific reads per cell to each multiplexed nucleus. Genome sequences for the mouse strains were downloaded from from NCBI `https://api.ncbi.nlm.nih.gov/datasets/genome/?taxon=10090` to generate a custom reference, and each genotype combination (B6J/NODJ, AJ/PWKJ, 129S1J/CASTJ, and WSBJ/NZOJ) was compared in each library. Nuclei were assigned a genotype based on the maximum number of counts in one of the two expected genotypes. Doublet detection was performed with Scrublet[97] in each sample within each subpool. Other than a baseline UMI

threshold of 200 UMIs per nucleus, the pipeline performs no additional filtering. Data from all nuclei belonging to the same tissue in the experiment were merged into a single H5ad file.

QC and clustering single-nucleus data

The count matrices and metadata as adata objects were further aggregated at the tissue level after all experiments were quantified. All tissues were filtered for nuclei with >500 and <150,000 UMIs, >250 expressed genes (>=1 UMI), <1% mitochondrial gene expression, and <0.25 doublet scores. Nuclei with ambiguous genotypes (around 0.2% of the dataset) were also excluded from downstream analysis. Scanpy[99] (version 1.9.5) was used to cluster nuclei for celltype and subtype annotation. Briefly, the counts matrices for each tissue were normalized by total UMI count followed by logarithmic transformation and filtration for highly variable genes (min_mean=0.0125, max_mean=3, min_disp=0.5). Percent mitochondrial gene expression and number of genes detected were regressed and normalized counts were scaled to unit variance and zero mean. Dimensionality reduction using PCA with the top 30 principal components was used to construct a neighborhood graph (n_neighbor = 20). Initial leiden clustering was performed for each tissue at the same resolution = 1. In most tissues (adrenal gland, gonads, diancephalon/pituitary, kidney, and liver), cell type annotation was performed in two rounds. The first round identified low-quality clusters that were discarded from the dataset. Clustering was performed again with the same parameters as above, except in kidney where maximum doublet score was decreased to 0.2.

Cell type annotation

Clustered nuclei were manually annotated using expression of known marker genes, discussions with expert collaborators, identification of cluster marker genes and cross-referencing with literature, and integration with label transfer from a reference dataset using scvi-tools when possible[347,348]. Cluster markers were determined by the rank_genes_groups Scanpy function with the t-test method. Annotation was performed at three levels, from fine to

coarse resolution: subtype, cell type, and general_cell type. For the most part, each cluster was assigned to a single subtype, with multiple clusters making up larger subtypes. Subclustering was performed in kidney, adrenal gland, diencephalon/pituitary, and male gonads datasets based on clear marker gene expression separating loop of Henle ascending and descending thin limb epithelial cells, adrenal cortical layers, corticotropes from thyrotropes, and mature sperm from spermatid. In liver, hepatocytes were annotated and separated from other cell types, which were clustered and annotated independently. Cell types are also annotated by a Cell Ontology (CL)[349] ID which matches the "cell type" level annotations. When necessary, "subtypes" extend CL cell types into cell subtypes and/or cell states. For example, "subtypes" captures specialized myonuclei such as those resting underneath the neuromuscular junction and myotendinous junction[112,113]. The "general_cell type" level uses the hierarchical structures embedded in the CL database to group cells into broaded annotations such as "epithelial cell" and "germ cell".

Pseudobulk differential expression analysis

Within a cell type such as satellite cells or proximal tubules, raw, unnormalized counts were extracted from the annotated Scanpy adata for each genotype pair (28 pairs total). The sample-level pseudobulk matrices grouped were calculated using the get_pseudobulk function from the decoupler package (mode='sum', min_cells=10, min_counts=10000). Multifactor differential expression analysis with pydeseq2[265] was used to compare genotypes to C57BL6/J and each other as well as females compared to males. Gene ontology analysis was performed with Metascape[350].

Calculating regulatory topics using Topyfic

Regulatory topics were calculated using a curated vocabulary of regulatory genes with the Topyfic package as previously described[128]. With a curated set of 2,789 regulatory genes and a chosen resolution k = 10 for all 9 tissues, we recovered between 7 and 13 topics. Since

satellite cells are a small population, we adjusted k to k = 8 and recovered 7 topics. All robust LDA topics were calculated using 100 runs. Structure plots were generated using the structure_plot function in Topyfic. Gene weights between topics of interest were compared for downstream analysis.

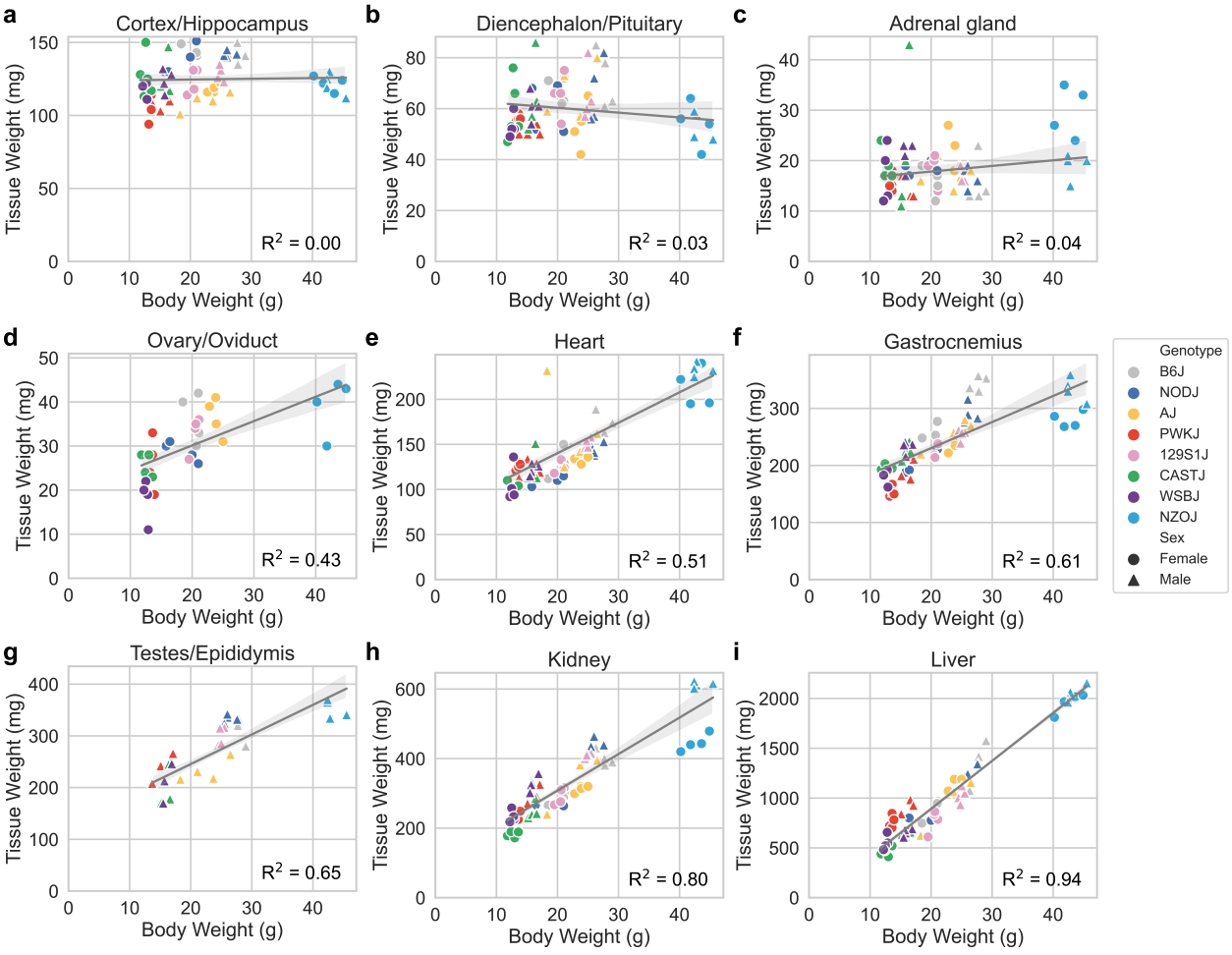

Figure 4.1: **Relationship between body weight and tissue weight in 9 diverse tissues a,** Body weight compared to cortex/hippocampus weight with a Pearson R^2 of 0.00. **b,** Body weight compared to diencephalon/pituitary weight with a Pearson R^2 of 0.03. **c,** Body weight compared to adrenal gland weight with a Pearson R^2 of 0.04. **d,** Body weight compared to ovary/oviduct weight with a Pearson R^2 of 0.43. **e,** Body weight compared to heart weight with a Pearson R^2 of 0.51. **f,** Body weight compared to gastrocnemius weight with a Pearson R^2 of 0.61. **g,** Body weight compared to testes/epididymis weight with a Pearson R^2 of 0.65. **h,** Body weight compared to kidney weight with a Pearson R^2 of 0.80. **i,** Body weight compared to liver weight with a Pearson R^2 of 0.94.

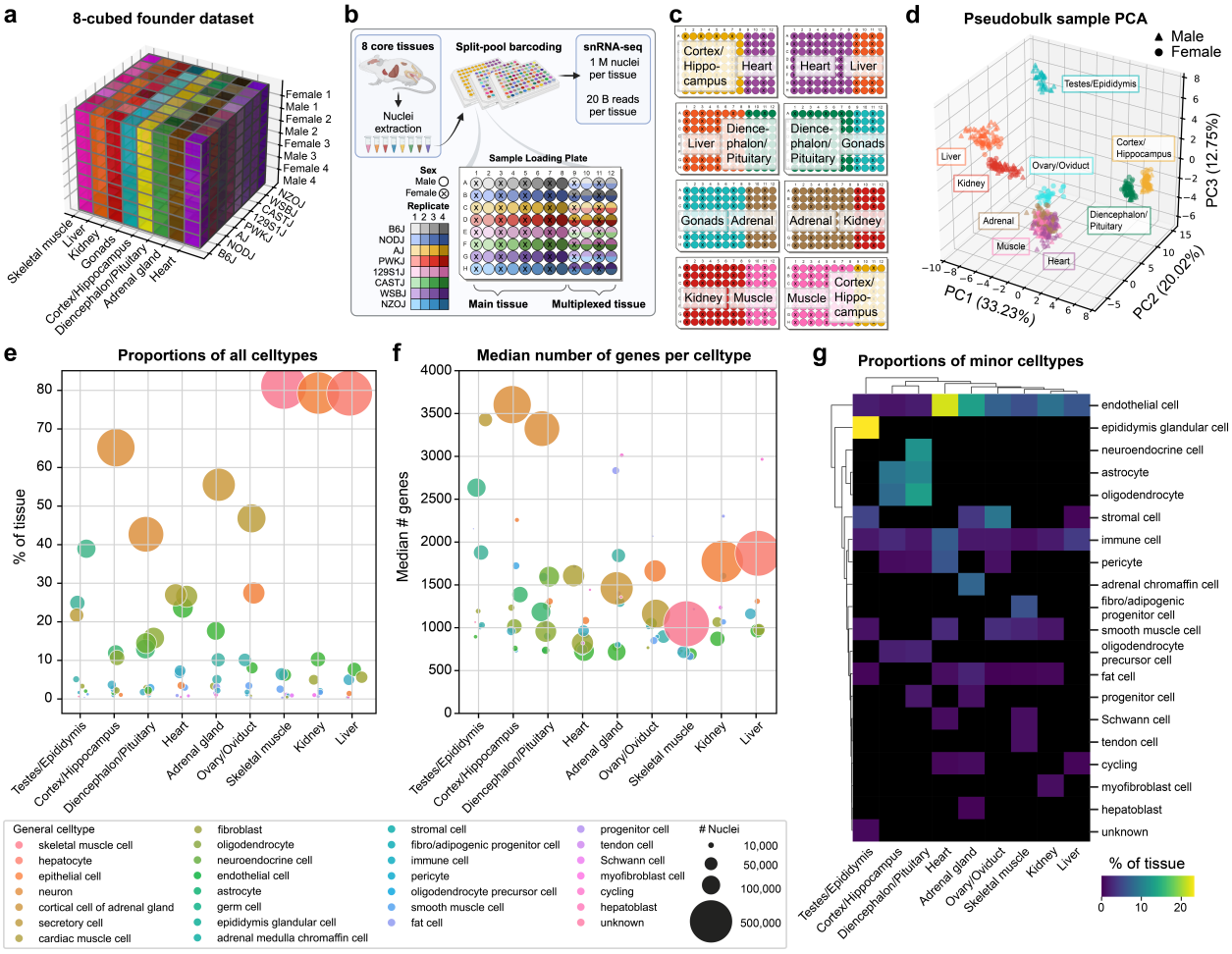

Figure 4.2: **Overview of the IGVF mouse dataset in 8 founder genotypes a,** Visualization of "8-cubed" dataset across 8 genotypes and 8 tissues/tissue groups replicated with 4 males and 4 females. **b,** Experimental design of sample barcoding plate. Nuclei derived from one tissue serves as the main tissue on the plate, where each sample has its own unique sample well and corresponding barcode. Nuclei from a different tissue are multiplexed in the remaining third of the plate, where two samples from distinct genotypes for the same sex and replicate are loaded into one well. **c,** Tissue loading pattern for all 8 sample barcoding plates. Note the first main tissue, cortex/hippocampus (yellow), serves as the multiplexed tissue in the final plate. **d,** Pseudobulk principal component analysis of 515 total samples colored by tissue. PCA was calculated using sum of raw counts across nuclei from individuals within each tissue with read depth normalization. **e,** Cell type proportions within each tissue, with point sizes reflecting the number of nuclei in each cell type. **f,** Median number of genes detected in nuclei within each cell type, with point sizes reflecting the number of nuclei in each cell type. **g,** Proportions of 20 minor cell types that make up less than 25% of each tissue.

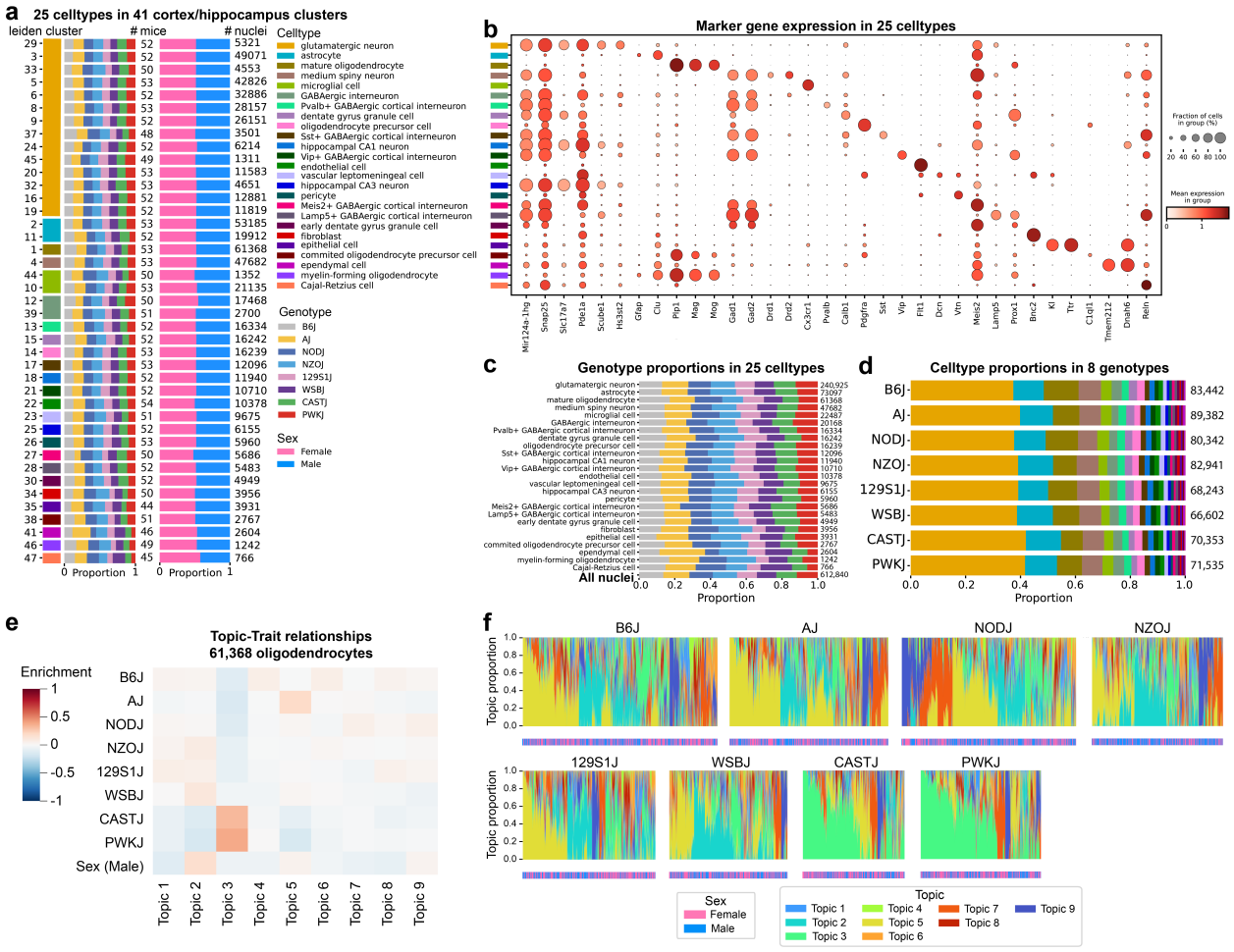

Figure 4.3: **Overview of celltypes recovered in cortex and hippocampus and regulatory topics modeling in oligodendrocytes a**, Proportion of genotype and sex in 41 cortex/hippocampus clusters. Clusters are ordered by largest to smallest annotated cell type (grouped color bars). Middle numeric column indicates the number of mice to constitute 90% of the cluster. Right numeric column indicates number of nuclei per cluster. **b**, Dot plot showing expression of marker genes in 25 annotated celltypes (refer to color legend in a). **c**, Distribution of genotypes in each cortex/hippocampus celltype. **d**, Distribution of celltypes in each genotype. **e**, Enrichment of genotype and sex in 9 regulatory topics. **f**, Structure plots of topic proportions per nucleus grouped by genotype. Each column is a stacked bar plot showing the proportion of participation across topics for each nucleus across 61,368 total nuclei in the celltype.

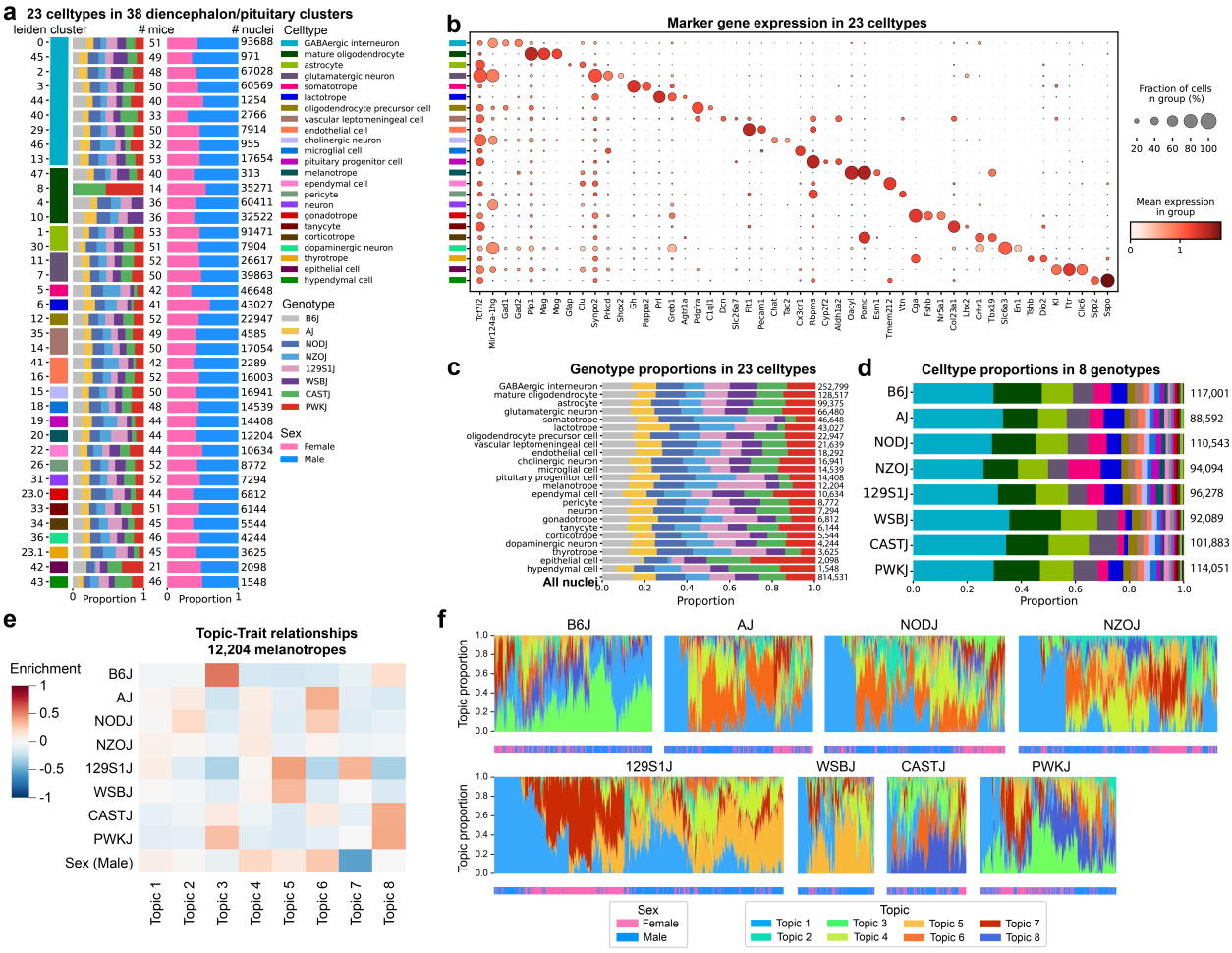

Figure 4.4: **Overview of celltypes recovered in diencephalon and pituitary gland and regulatory topics modeling in melanotropes a**, Proportion of genotype and sex in 38 diencephalon/pituitary clusters. Clusters are ordered by largest to smallest annotated cell type (grouped color bars). Middle numeric column indicates the number of mice to constitute 90% of the cluster. Right numeric column indicates number of nuclei per cluster. **b**, Dot plot showing expression of marker genes in 23 annotated celltypes (refer to color legend in a). **c**, Distribution of genotypes in each diencephalon/pituitary celltype. **d**, Distribution of celltypes in each genotype. **e**, Enrichment of genotype and sex in 8 regulatory topics. **f**, Structure plots of topic proportions per nucleus grouped by genotype. Each column is a stacked bar plot showing the proportion of participation across topics for each nucleus across 12,204 total nuclei in the celltype.

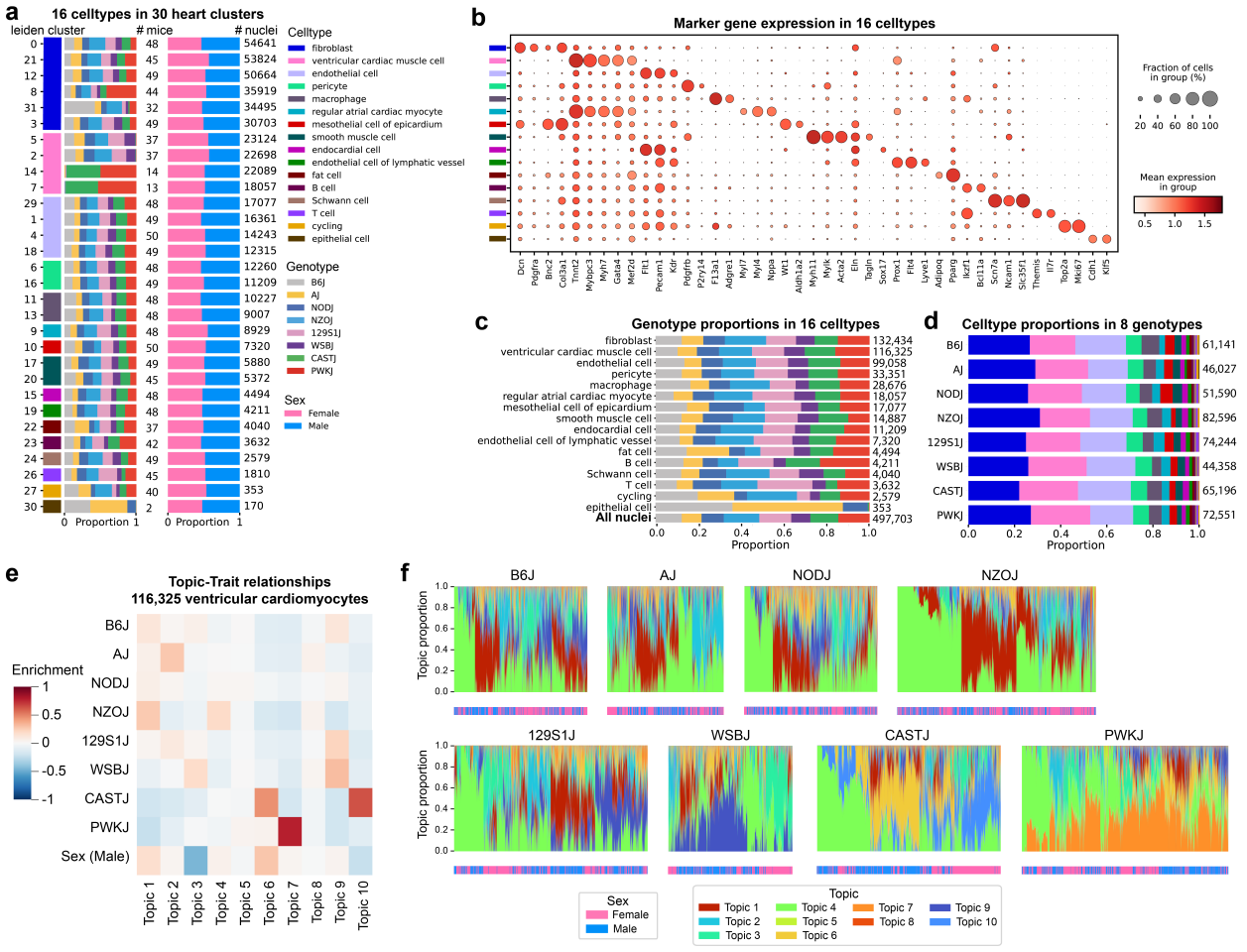

Figure 4.5: **Overview of celltypes recovered in heart and regulatory topics modeling in ventricular cardiomyocytes a**, Proportion of genotype and sex in 30 heart clusters. Clusters are ordered by largest to smallest annotated cell type (grouped color bars). Middle numeric column indicates the number of mice to constitute 90% of the cluster. Right numeric column indicates number of nuclei per cluster. **b**, Dot plot showing expression of marker genes in 16 annotated celltypes (refer to color legend in a). **c**, Distribution of genotypes in each heart celltype. **d**, Distribution of celltypes in each genotype. **e**, Enrichment of genotype and sex in 10 regulatory topics. **f**, Structure plots of topic proportions per nucleus grouped by genotype. Each column is a stacked bar plot showing the proportion of participation across topics for each nucleus across 116,325 total nuclei in the celltype.

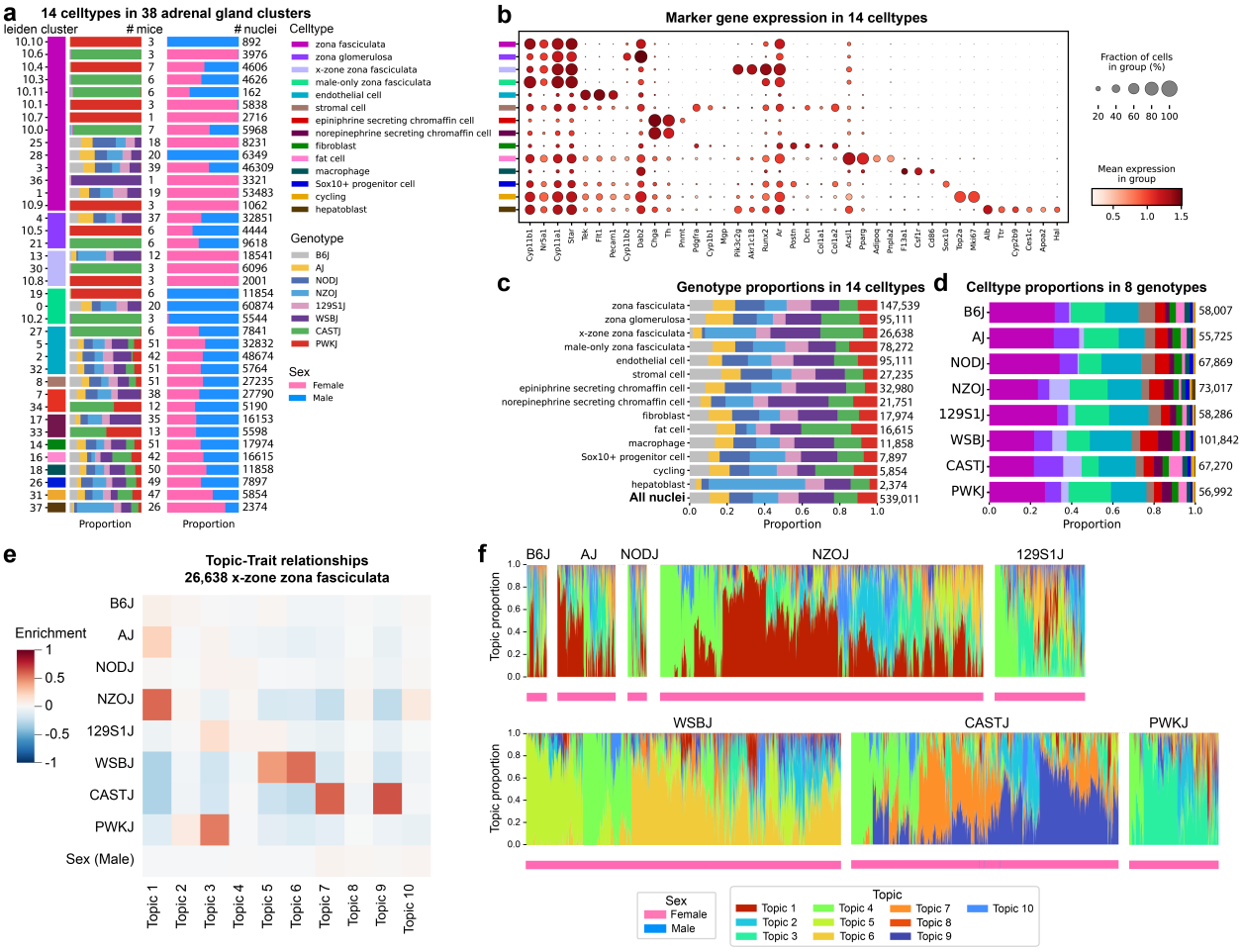

Figure 4.6: **Overview of celltypes recovered in adrenal gland and regulatory topics modeling in X-zone a**, Proportion of genotype and sex in 38 adrenal clusters. Clusters are ordered by largest to smallest annotated cell type (grouped color bars). Middle numeric column indicates the number of mice to constitute 90% of the cluster. Right numeric column indicates number of nuclei per cluster. **b**, Dot plot showing expression of marker genes in 14 annotated celltypes (refer to color legend in a). **c**, Distribution of genotypes in each adrenal celltype. **d**, Distribution of celltypes in each genotype. **e**, Enrichment of genotype and sex in 10 regulatory topics. **f**, Structure plots of topic proportions per nucleus grouped by genotype. Each column is a stacked bar plot showing the proportion of participation across topics for each nucleus across 26,606 total nuclei in the celltype.

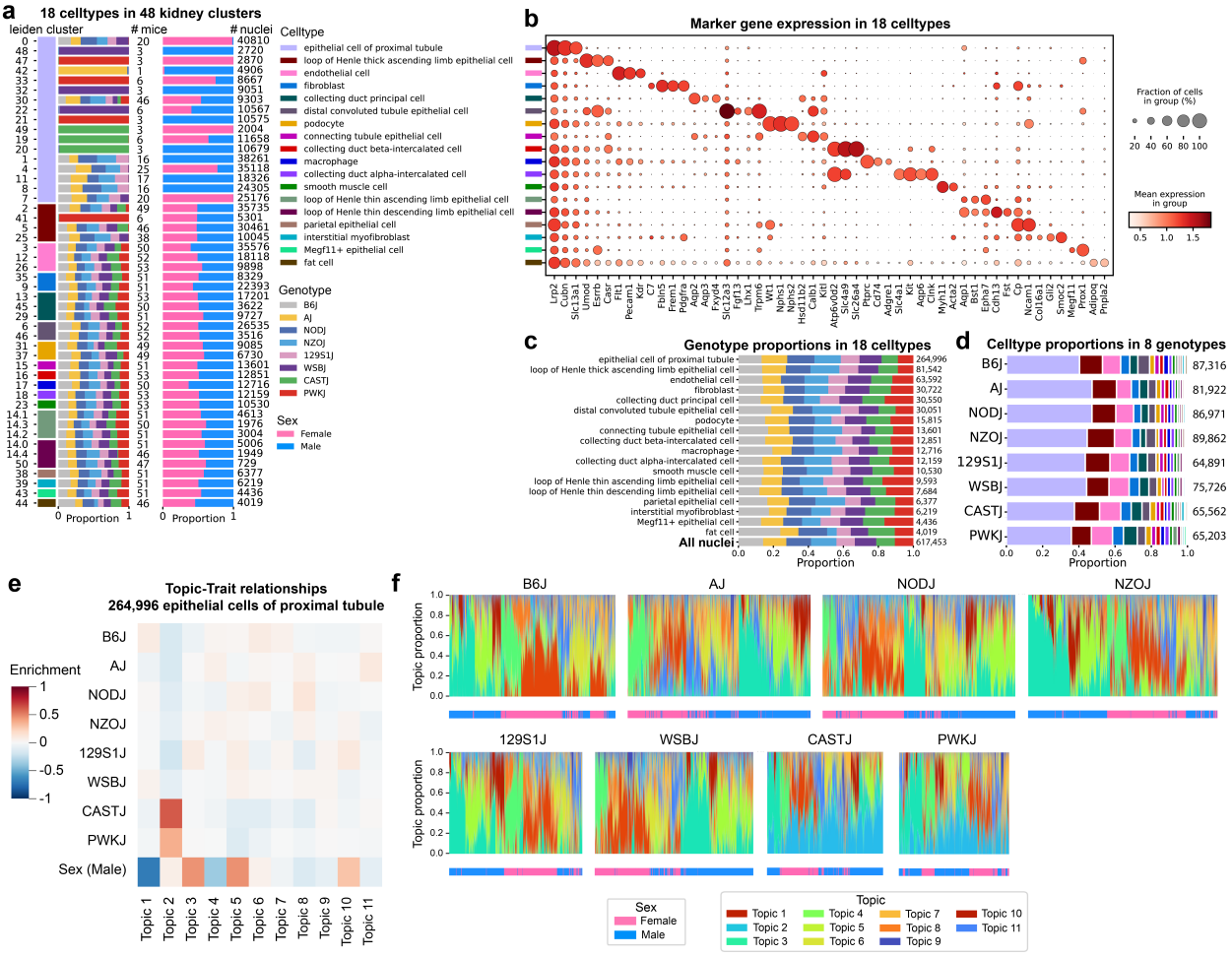

Figure 4.7: **Overview of celltypes recovered in kidney and regulatory topics modeling in proximal tubules a**, Proportion of genotype and sex in 48 kidney clusters. Clusters are ordered by largest to smallest annotated cell type (grouped color bars). Middle numeric column indicates the number of mice to constitute 90% of the cluster. Right numeric column indicates number of nuclei per cluster. **b**, Dot plot showing expression of marker genes in 18 annotated celltypes (refer to color legend in a). **c**, Distribution of genotypes in each kidney celltype. **d**, Distribution of celltypes in each genotype. **e**, Enrichment of genotype and sex in 11 regulatory topics. **f**, Structure plots of topic proportions per nucleus grouped by genotype. Each column is a stacked bar plot showing the proportion of participation across topics for each nucleus across 264,996 total nuclei in the celltype.

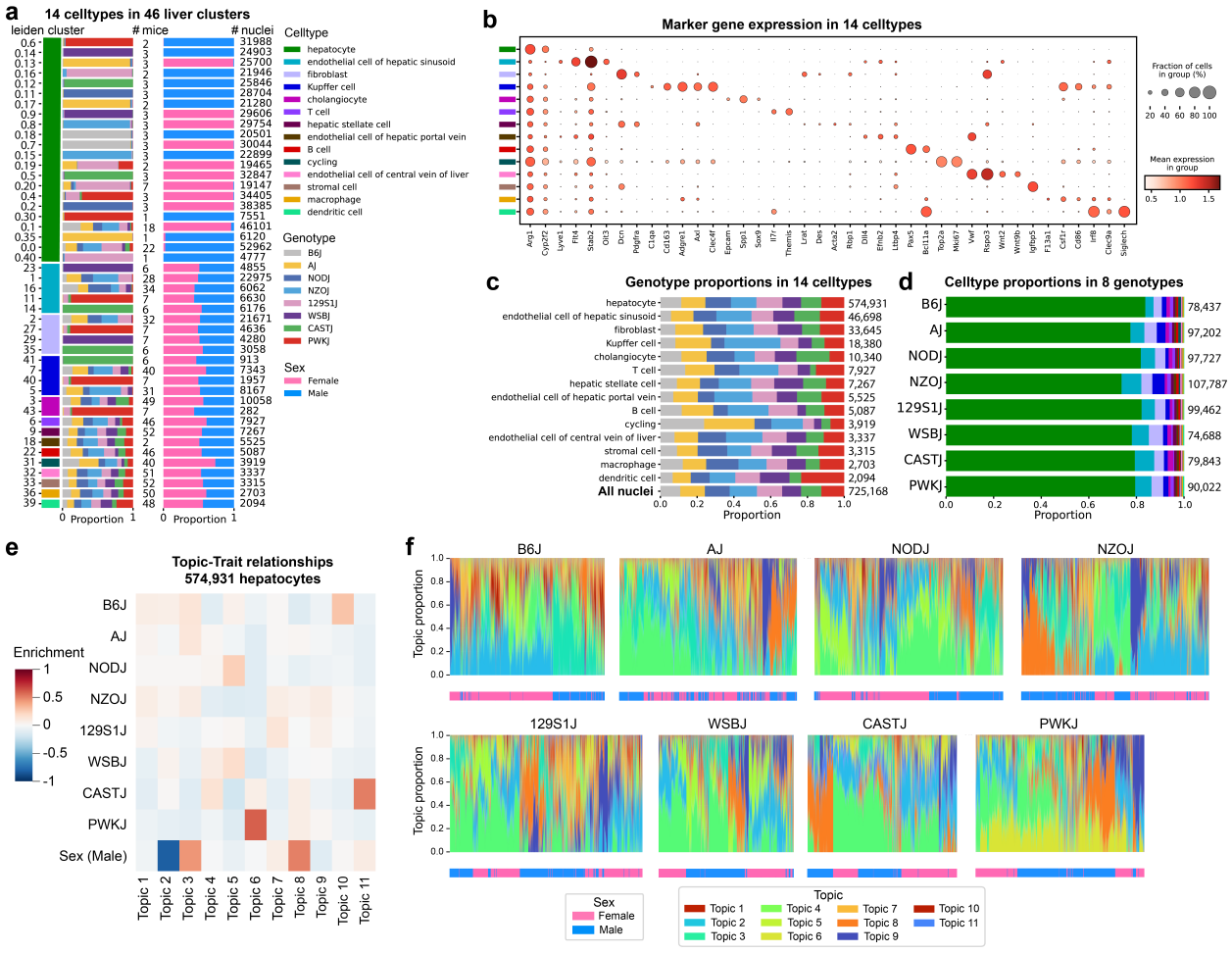

Figure 4.8: **Overview of celltypes recovered in liver and regulatory topics modeling in hepatocytes a**, Proportion of genotype and sex in 48 liver clusters. Clusters are ordered by largest to smallest annotated cell type (grouped color bars). Middle numeric column indicates the number of mice to constitute 90% of the cluster. Right numeric column indicates number of nuclei per cluster. **b**, Dot plot showing expression of marker genes in 14 annotated celltypes (refer to color legend in a). **c**, Distribution of genotypes in each liver celltype. **d**, Distribution of celltypes in each genotype. **e**, Enrichment of genotype and sex in 11 regulatory topics. **f**, Structure plots of topic proportions per nucleus grouped by genotype. Each column is a stacked bar plot showing the proportion of participation across topics for each nucleus across 574,931 total nuclei in the celltype.

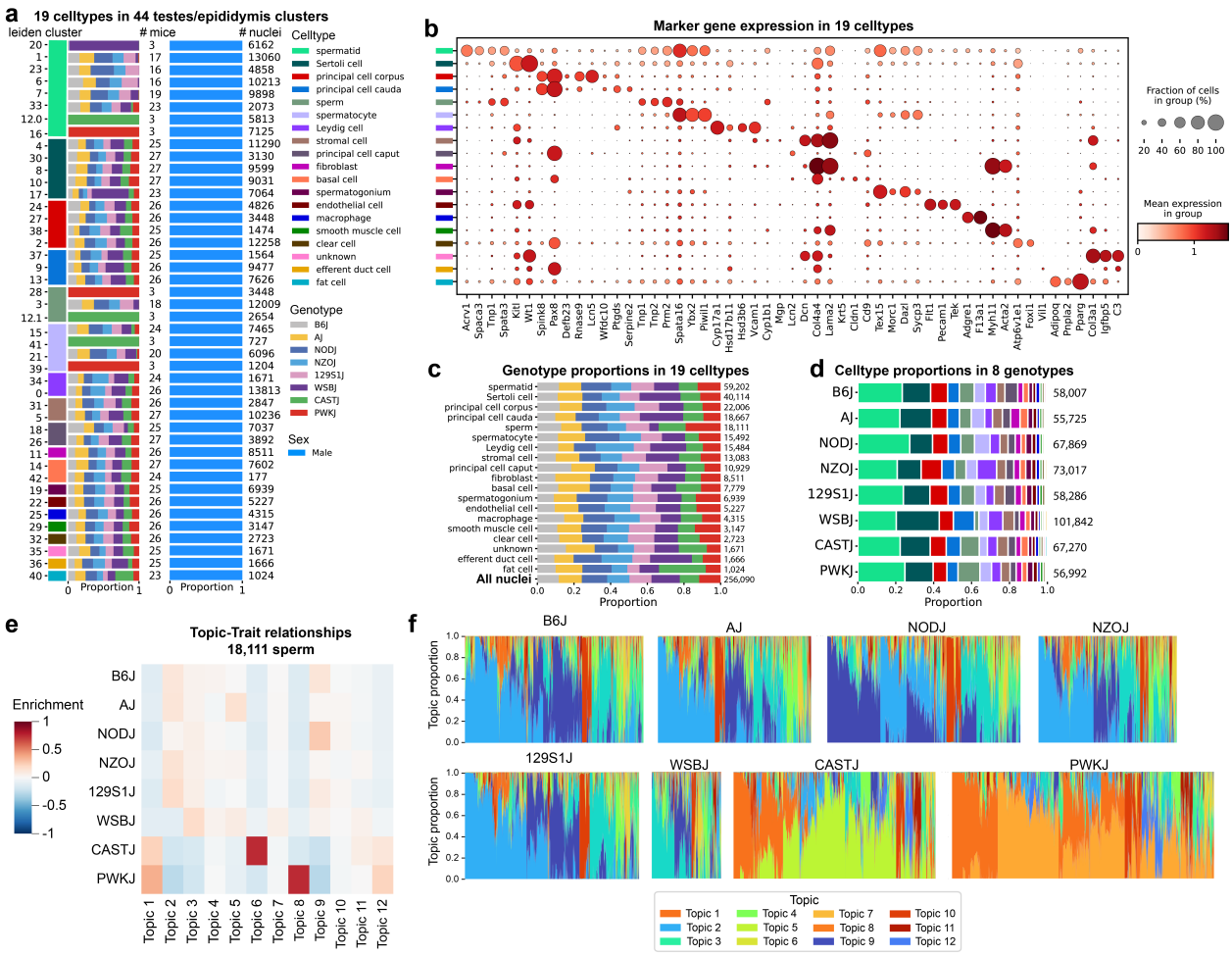

Figure 4.9: **Overview of celltypes recovered in testes and epididymis and regulatory topics modeling in sperm cells a**, Proportion of genotype and sex in 44 testes/epididymis clusters. Clusters are ordered by largest to smallest annotated cell type (grouped color bars). Middle numeric column indicates the number of mice to constitute 90% of the cluster. Right numeric column indicates number of nuclei per cluster. **b**, Dot plot showing expression of marker genes in 19 annotated celltypes (refer to color legend in a). **c**, Distribution of genotypes in each testes/epididymis celltype. **d**, Distribution of celltypes in each genotype. **e**, Enrichment of genotype and sex in 12 regulatory topics. **f**, Structure plots of topic proportions per nucleus grouped by genotype. Each column is a stacked bar plot showing the proportion of participation across topics for each nucleus across 18,111 total nuclei in the celltype.

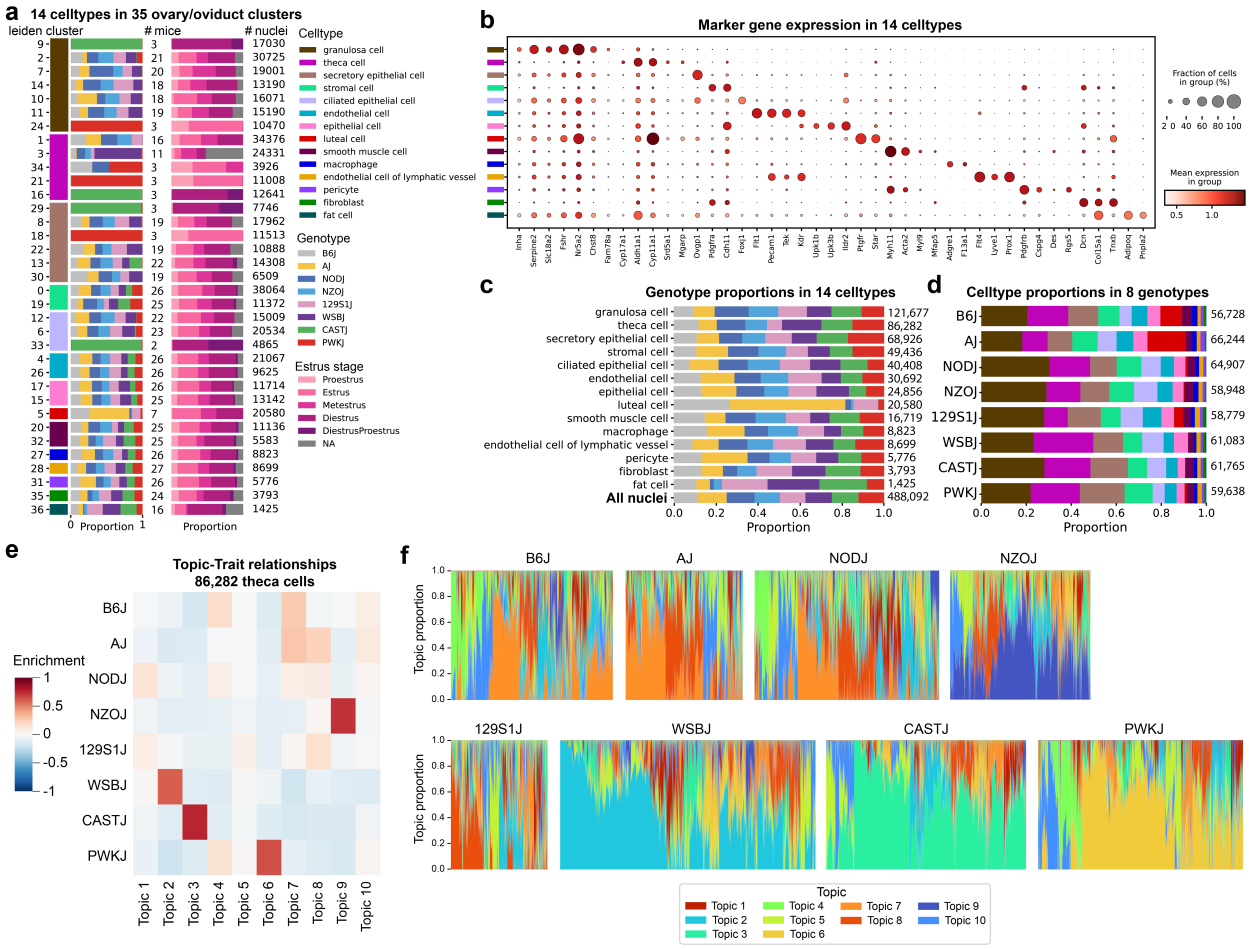

Figure 4.10: **Overview of celltypes recovered in ovary and oviduct and regulatory topics modeling in theca cells a**, Proportion of genotype and sex in 35 ovary/oviduct clusters. Clusters are ordered by largest to smallest annotated cell type (grouped color bars). Middle numeric column indicates the number of mice to constitute 90% of the cluster. Right numeric column indicates number of nuclei per cluster. Right bar plot indicates estrus stage. **b**, Dot plot showing expression of marker genes in 14 annotated celltypes (refer to color legend in a). **c**, Distribution of genotypes in each ovary/oviduct celltype. **d**, Distribution of celltypes in each genotype. **e**, Enrichment of genotype and sex in 10 regulatory topics. **f**, Structure plots of topic proportions per nucleus grouped by genotype. Each column is a stacked bar plot showing the proportion of participation across topics for each nucleus across 86,282 total nuclei in the celltype.

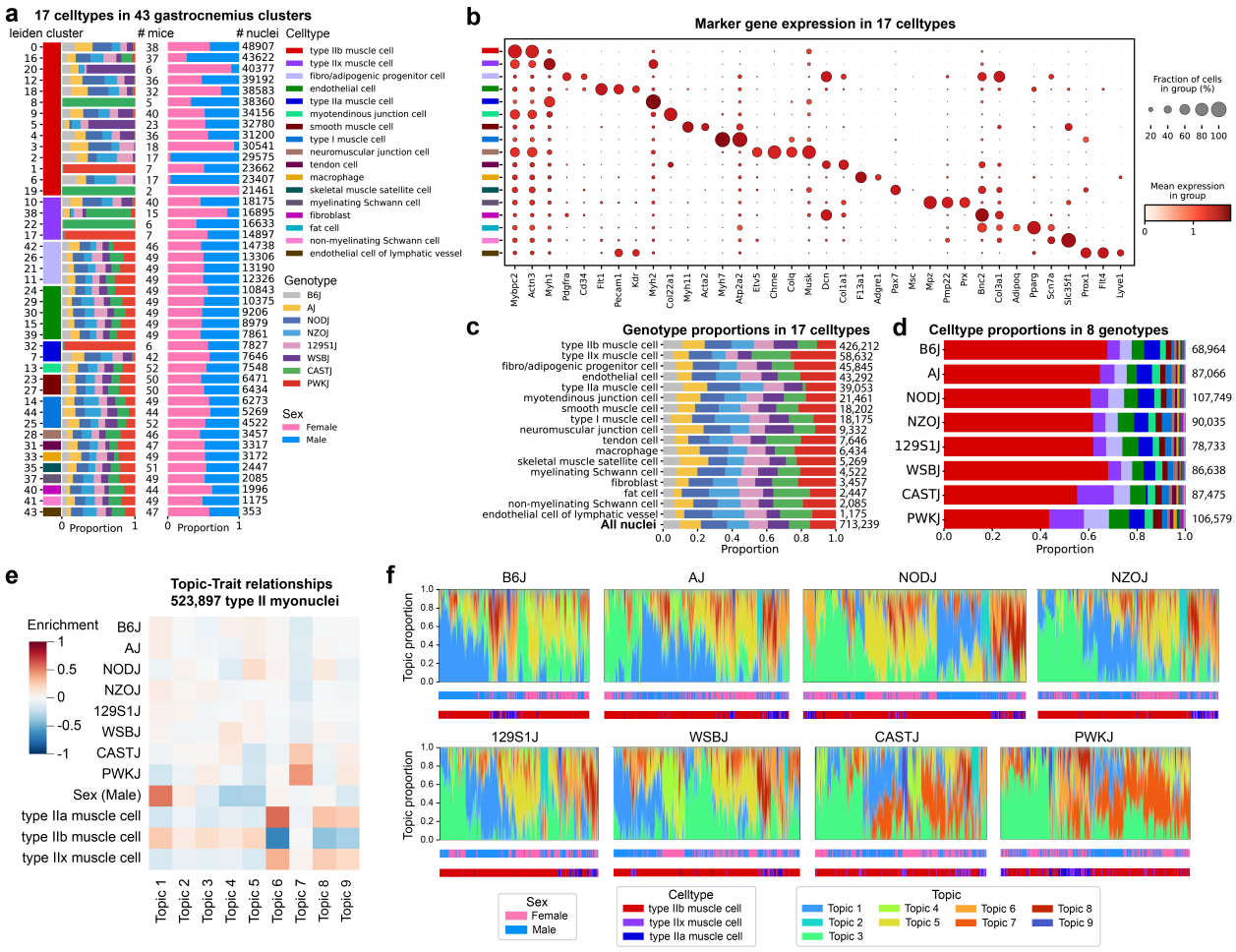

Figure 4.11: **Overview of celltypes recovered in skeletal muscle and regulatory topics modeling in type II myonuclei a**, Proportion of genotype and sex in 43 skeletal muscle clusters. Clusters are ordered by largest to smallest annotated cell type (grouped color bars). Middle numeric column indicates the number of mice to constitute 90% of the cluster. Right numeric column indicates number of nuclei per cluster. **b**, Dot plot showing expression of marker genes in 17 annotated celltypes (refer to color legend in a). **c**, Distribution of genotypes in each skeletal muscle celltype. **d**, Distribution of celltypes in each genotype. **e**, Enrichment of genotype and sex in 11 regulatory topics. **f**, Structure plots of topic proportions per nucleus grouped by genotype. Each column is a stacked bar plot showing the proportion of participation across topics for each nucleus across 523,897 total nuclei in type IIb myonuclei.

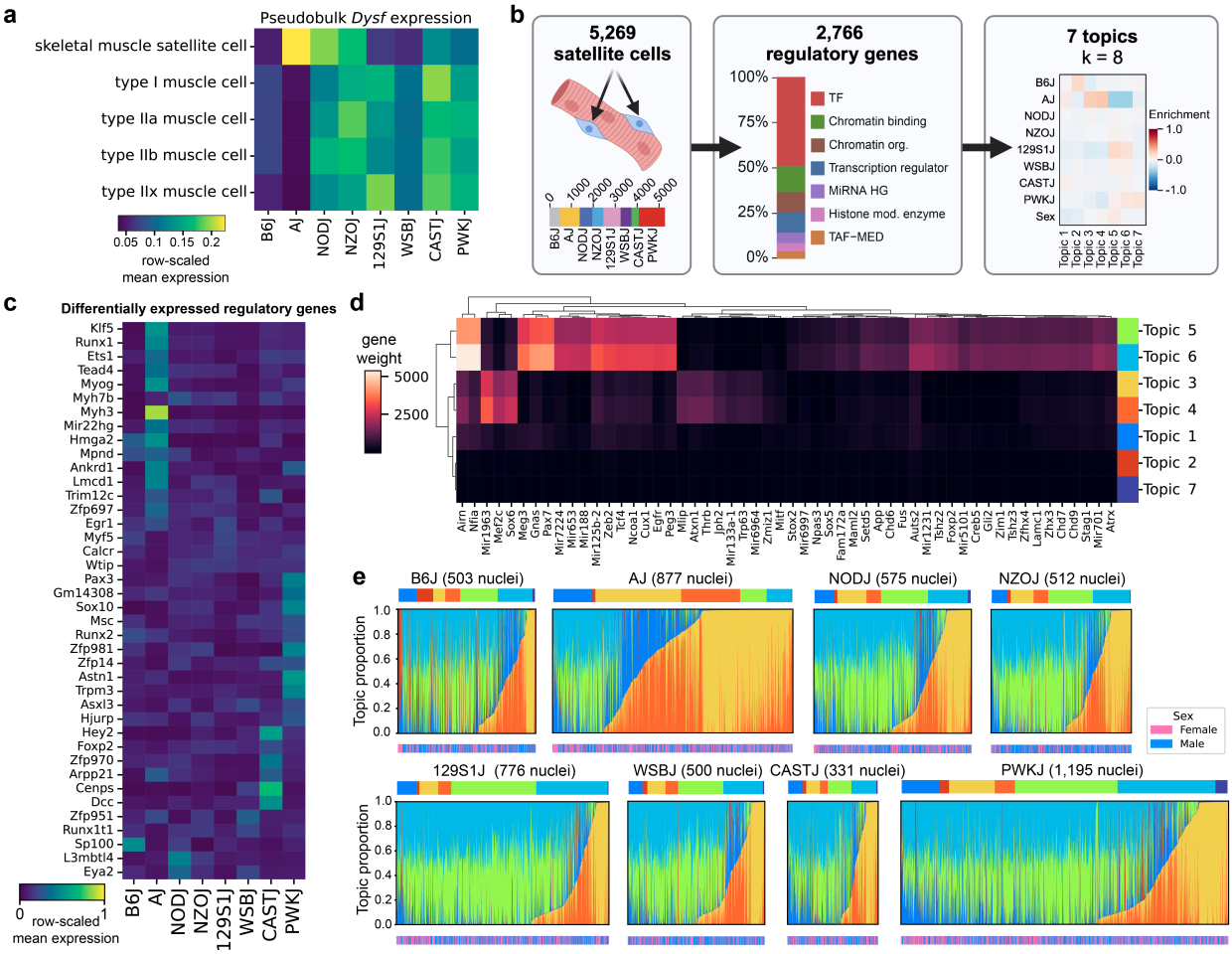

Figure 4.12: **Satellite cells are activated in AJ a**, Heatmap of pseudobulk dysferlin expression in satellite cells (muscle stem cells) and myonuclei subtypes grouped by genotype. **b**, Overview of topics modeling analysis on 5,269 satellite cells, similar to the approach used for specific celltypes in all tissues. **c** Differentially expressed genes (log fold change > 2.5 and adjusted p-value < 0.01) upregulated in one specific genotype overlapping with regulatory genes. **d**, Heatmap of regulatory genes differentially weighted between AJ-specific topics 3 and 4 compared to shared topics 5 and 6. **e,** Structure plots of topic proportions per nucleus grouped by genotype. Each column is a stacked bar plot showing the proportion of participation across topics for each nucleus across 5,269 total nuclei in satellite cells.

Chapter 5

Future directions

Towards cohesive cell type annotation

As single-cell technologies become more mainstream, initial "atlases" of cell types are built in a wide range of tissues and conditions. These studies typically profile whole transcriptomes or genome-wide chromatin accessibility in increasingly large numbers of single cells[196,198,351,352]. While some studies organize cell types into hierarchical trees[352] or taxonomies[196], most represent them as discrete clusters within a single tissue. Cell types are described using human-readable terms[198,351] or more obscure acronyms[196]. Integrated analysis involves relabeling cell types with a unified naming strategy to address inconsistencies across studies[352]. Thus, there is a pressing need for a widely agreed-upon and cohesive naming convention in the single-cell field that remains flexible to different levels of resolution. Such a convention should accommodate increasingly rarer but distinct cell types and states while preserving biological relationships between previously annotated cell types. Initiatives such as the Cell Ontology (CL) offer a structured approach, providing human-readable descriptors, synonyms, numerical IDs, and hierarchical structures for cell types[349]. However, CL currently does not encompass distinct cell states within individual nuclei such as those observed in specialized

myonuclei[112]. Expanding CL to include such states would significantly benefit the single-cell field as a whole.

Long-read single-cell RNA-seq

While standard short-read single-cell RNA-seq is the most widely used method for quantifying transcript expression in single cells, it falls short in preserving the structure of alternatively spliced full-length isoforms. Alternative splicing is a co-transcriptional regulatory process[353] occurring in almost every mammalian gene that enhances the diversity of the proteome. Transcript isoforms encoding distinct proteins may possess different functional properties that contribute to cell type specialization. For instance, the length of isoforms for the sarcomere protein titin correlates with muscle fiber extensibility[354]. Shorter isoforms expressed solely in cardiac tissue yield proteins with higher passive tensile strength than those encoded by longer isoforms[354]. Long-read RNA-seq has recently been adapted to single-cell platforms to identify isoforms expressed in heterogeneous populations but is hindered by either low throughput, limiting resolution for discerning less abundant cell types, or high expense[131,133]. We addressed this issue in our 2021 paper by employing combinatorial barcoding to subset cells or nuclei from the main barcoded pool and sequence them with both long and short reads using a dual library preparation strategy from the same input cDNA. The remaining majority of cells or nuclei are deeply sequenced with more cost-effective short reads. However, our reliance on single nuclei extracted from flash-frozen tissues instead of whole cells introduced downstream artifacts evident during long-read data processing, particularly the capture of unspliced RNA from the nucleus. These artifacts lead to non-full length reads and negatively impact isoform detection. Performing exon capture enrichment on the barcoded cDNA before long-read sequencing substantially improves the fraction of fully spliced reads, enhancing isoform quantification and facilitating the discovery of novel isoforms[355].

Leveraging F1 crosses to compare gene regulation in *cis* and *trans*

Crossing distinct strains such as C57BL/6J and CAST/EiJ produces the first filial generation or F1 (B6CASTF1/J). The F1 offspring inherits one set of alleles from each homozygous parent. These crosses allow for investigation of the genetic basis of traits by comparing molecular characteristics in F1 hybrids with those of the parental strains. The F1s may exhibit similar, diminished, or augmented biological qualities compared to either parent. For example, B6CASTF1/J males grow beyond the size of either parent in terms of body weight[356]. The impact of a particular allele can be determined as either *cis* or *trans* by examining the expression levels of a specific transcript or protein in both the parents and offspring[357,358]. In the case of a *cis* effect, the allelic expression from one parent relative to the allelic expression from the other parent in the F1 hybrids mirrors the expression ratio observed when comparing the homozygous parents. In contrast, genes undergoing *trans* regulation have alleles that are equally expressed in the F1 hybrids. This indicates that *trans*-regulatory factors interact with target sequences to regulate both alleles regardless of the parental origin. As an example, the allele for age-related hearing loss noted in C57BL/6J was mapped to a locus on chromosome 10 using B6CASTF1/J hybrids and back-crossing in 1997 (before the gene itself was even identified[359]). While C57BL/6J suffer from hearing loss within a year after birth, CAST/EiJ and B6CASTF1/J have good hearing until at least 18 months old. The variation was later mapped as a *cis*-acting SNP in *Cdh23*[260].

Combining the founder snRNA-seq data we generated in IGVF and the B6CASTF1/J snRNA-seq data we produced in ENCODE4 will allow us to begin testing and developing allele-specific gene expression pipelines to map *cis* and *trans* regulatory effects at the cell type level. In addition, the second phase of our IGVF project includes snRNA-seq of our eight core tissues in F1 mice (C57BL/6J females crossed with each of the seven founder strains). This additional data in all eight core tissues will provide further opportunity for allele-specific gene expression analysis and determination of *cis* and *trans* regulatory effects in a broad spectrum of cell types and states.

Mapping cell type-specific QTLs

The CC lines have been used to map gene expression QTLs (eQTLs) and chromatin accessibility QTLs (caQTLs) in diverse tissues[65]. Transcript expression was captured through bulk RNA-seq while bulk ATAC-seq identified open chromatin across whole tissues in 47 CC strains. Flanking marker loci were used to infer haplotype blocks and the probability of each founder strain being the ancestor of a given allele was calculated[65,360]. A linear model was used to test the genetic effect at each locus as described by Keele et al., 2020[65]:

$$y_i = \mu + \text{batch}_{b[i]} + \text{QTL}_i + \varepsilon$$

Here, y_i represents the trait level for each individual i, μ is the intercept, $\text{batch}_{b[i]}$ accounts for sequencing batches, ε denotes the residual noise, and QTL_i is defined as $\text{QTL}_i = \beta^T x_i$. In this equation, $x_i = (x_{i,\text{AJ}}, ..., x_{i,\text{WSB}})^T$ is a vector of inferred haplotype dosages for the eight founders in each individual and $\beta = (\beta_{\text{AJ}}, ..., \beta_{\text{WSB}})^T$ corresponds to the additive effect of each haplotype. The fit of the linear model to the RNA-seq and ATAC-seq data is compared with and without the QTL term to obtain a p-value. Significant QTLs were identified for both genes and chromatin regions to highlight local QTLs with *cis* effects and distal QTLs with *trans* effects. Certain eQTLs and caQTLs were found to be shared across multiple tissues, while others exhibited tissue-specificity[65]. Tissue-specific effects can be attributed to distinct regulatory contexts and unique cell type compositions shaped by gene regulatory programs. Genes within these programs may interact with different sets of transcription factors and regulatory elements across tissues to contribute to tissue-specific eQTL effects. This perspective extends to the resolution of individual cell types which also exhibit extensive specificity in gene expression

While QTLs have traditionally been identified in bulk tissues using the CC panel, our focus with single-nucleus RNA-seq dataset is to map eQTLs at the level of specific cell types[361].

We aim to collect snRNA-seq data across our eight core tissues from 33 CC lines with further plans to include snATAC-seq data in some or all matching samples. The addition of snATAC-seq data in particular enhances QTL resolution, since CC haplotypes are often large and span tens of megabases. Chromatin accessibility within these regions can help pinpoint the genomic locus of active QTLs. The insights gained from mining single-cell-level QTLs from this large-scale dataset will be invaluable to the mouse research community, particularly those utilizing the Collaborative Cross and Diversity Outbred panels as well as human systems geneticists working with more complex genomes.

Bibliography

[1] Alfred H. Sturtevant. The Early Mendelians. *Genetics*, 109(4):199–204, 1965. ISSN 1943-2631.

[2] Michael Turelli. Fisher's infinitesimal model: A story for the ages. *Theoretical Population Biology*, 118:46–49, 2017. ISSN 0040-5809. doi: 10.1016/j.tpb.2017.09.003.

[3] F J Ayala and W M. Fitch. Genetics and the origin of species: an introduction. *Proceedings of the National Academy of Sciences of the United States of America*, 94 (15):7691–7697, 1997. ISSN 1091-6490. doi: 10.1073/pnas.94.15.7691.

[4] J. D. Watson and F. H. Crick. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(5156):737–738, 1953. ISSN 0028-0836. doi: 10.1038/171737a0.

[5] A. Klug. Rosalind Franklin and the discovery of the structure of DNA. *Nature*, 219 (5156):808–810, 1968. ISSN 0028-0836. doi: 10.1038/219808a0.

[6] Asude Alpman Durmaz, Emin Karaca, Urszula Demkow, Gokce Toruner, Jacqueline Schoumans, and Ozgur Cogulu. Evolution of genetic techniques: past, present, and beyond. *BioMed Research International*, 2015:461524, 2015. ISSN 2314-6141. doi: 10.1155/2015/461524.

[7] C. Miles and M Wayne. Quantitative Trait Locus (QTL) Analysis. *Nature Education*, 1(1):208, 2008.

[8] K Weber, R Eisman, L Morey, A Patty, J Sparks, M Tausek, and Z B. Zeng. An analysis of polygenes affecting wing shape on chromosome 3 in Drosophila melanogaster. *Genetics*, 153(2):773–786, 1999. ISSN 1943-2631. doi: 10.1093/genetics/153.2.773.

[9] M C Gurganus, J D Fry, S V Nuzhdin, E G Pasyukova, R F Lyman, and T F. Mackay. Genotype-environment interaction at quantitative trait loci affecting sensory bristle number in Drosophila melanogaster. *Genetics*, 149(4):1883–1898, 1998. ISSN 1943-2631. doi: 10.1093/genetics/149.4.1883.

[10] T. F. Mackay. Quantitative trait loci in Drosophila. *Nature Reviews Genetics*, 2(1): 11–20, 2001. ISSN 1471-0064. doi: 10.1038/35047544.

[11] Sergey V Nuzhdin, Aziz A Khazaeli, and James W. Curtsinger. Survival analysis of life span quantitative trait loci in Drosophila melanogaster. *Genetics*, 170(2):719–731, 2005. ISSN 1943-2631. doi: 10.1534/genetics.104.038331.

[12] Jeff Leips and Trudy F C Mackay. The complex genetic architecture of Drosophila life span. *Experimental Aging Research*, 28(4):361–90, 2002. ISSN 1096-4657. doi: 10.1080/03610730290080399.

[13] Scott N Forbes, Robert K Valenzuela, Paul Keim, and Philip M. Service. Quantitative trait loci affecting life span in replicated populations of Drosophila melanogaster. I. Composite interval mapping. *Genetics*, 168(1):301–311, 2004. ISSN 1943-2631. doi: 10.1534/genetics.103.023218.

[14] H. Forsberg and P. O. Ljungdahl. Sensors of extracellular nutrients in Saccharomyces cerevisiae. *Current Genetics*, 40(2):91–109, 2001. ISSN 1432-0983. doi: 10.1007/s002940100244.

[15] Lars M. Steinmetz, Himanshu Sinha, Dan R. Richards, Jamie I. Spiegelman, Peter J. Oefner, John H. McCusker, and Ronald W. Davis. Dissecting the architecture of a quantitative trait locus in yeast. *Nature*, 416(6878):326–330, 2002. ISSN 0028-0836.

doi: 10.1038/416326a.

[16] Rachel B Brem and Leonid Kruglyak. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, 102(5):1572–1577, 2005. ISSN 1091-6490. doi: 10.1073/pnas.0408709102.

[17] Michael P Snyder, Thomas R Gingeras, Jill E Moore, Zhiping Weng, Mark B Gerstein, Bing Ren, Ross C Hardison, John A Stamatoyannopoulos, Brenton R Graveley, Elise A Feingold, Michael J Pazin, Michael Pagan, Daniel A Gilchrist, Benjamin C Hitz, J Michael Cherry, Bradley E Bernstein, Eric M Mendenhall, Daniel R Zerbino, Adam Frankish, Paul Flicek, and Richard M Myers. Perspectives on ENCODE. *Nature*, 583(7818):693–698, 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2449-8.

[18] IGVF Consortium. The Impact of Genomic Variation on Function (IGVF) Consortium. *ArXiv*, page 2307.13708v1, 2023.

[19] Alessandra Breschi, Thomas R. Gingeras, and Roderic Guigó. Comparative transcriptomics in human and mouse. *Nature Reviews Genetics*, 18(7):425–440, 2017. ISSN 1471-0064. doi: 10.1038/nrg.2017.19.

[20] Toyoyuki Takada, Toshinobu Ebata, Hideki Noguchi, Thomas M Keane, David J Adams, Takanori Narita, Tadasu Shin-I, Hironori Fujisawa, Atsushi Toyoda, Kuniya Abe, Yuichi Obata, Yoshiyuki Sakaki, Kazuo Moriwaki, Asao Fujiyama, Yuji Kohara, and Toshihiko Shiroishi. The ancestor of extant Japanese fancy mice contributed to the mosaic genomes of classical inbred strains. *Genome Research*, 23(8):1329–1338, 2013. ISSN 1088-9051. doi: 10.1101/gr.156497.113.

[21] Kenneth Paigen. One hundred years of mouse genetics: an intellectual history. I. The classical period (1902-1980). *Genetics*, 163(1):1–7, 2003. ISSN 1943-2631. doi: 10.1093/genetics/163.1.1.

[22] William E. Castle. The laws of Galton and Mendel and some laws governing race improvement by selection. *Proceedings of the American Academy of Arts and Sciences*, 39(8), 1903. ISSN 0199-9818.

[23] James F. Crow. C. C. Little, cancer and inbred mice. *Genetics*, 161(4):1357–1361, 2002. ISSN 1943-2631. doi: 10.1093/genetics/161.4.1357.

[24] C. Morse III Herbert. *Origins of inbred mice*. Academic Press, 1978. ISBN 0124123686.

[25] The Jackson Laboratory. The jackson laboratory homepage, 2024. URL `https://www.jax.org/`.

[26] David P. Steensma, Robert A. Kyle, and Marc A. Shampo. Abbie Lathrop, the "Mouse Woman of Granby": Rodent Fancier and Accidental Genetics Pioneer. *Mayo Clinic Proceedings*, 85(11):e83, 2010. ISSN 1942-5546. doi: 10.4065/mcp.2010.0647.

[27] Karen Artzt. Mammalian developmental genetics in the twentieth century. *Genetics*, 192(4):1151–1163, 2012. ISSN 1943-2631. doi: 10.1534/genetics.112.146191.

[28] R. B. Jackson and C. C. Little. The existence of non-chromosomal influence in the incidence of mammary tumors in mice. *Science*, 78(2029):465–466, 1933. ISSN 1095-9203. doi: 10.1126/science.78.2029.465.

[29] Camron D. Bryant. The blessings and curses of C57BL/6 substrains in mouse genetic studies. *Annals of the New York Academy of Sciences*, (1245):31–33, 2016. ISSN 1749-6632. doi: 10.1111/j.1749-6632.2011.06325.x.

[30] Mouse Genome Sequencing Consortium, Robert H. Waterston, Kerstin Lindblad-Toh,

Ewan Birney, Jane Rogers, Josep F Abril, Pankaj Agarwal, Richa Agarwala, Rachel Ainscough, Marina Alexandersson, Peter An, Stylianos E. Antonarakis, John Attwood, Robert Baertsch, Jonathon Bailey, Karen Barlow, Stephan Beck, Eric Berry, Bruce Birren, Toby Bloom, Peer Bork, Marc Botcherby, Nicolas Bray, Michael R. Brent, Daniel G. Brown, Stephen D. Brown, Carol Bult, John Burton, Jonathan Butler, Robert D. Campbell, Piero Carninci, Simon Cawley, Francesca Chiaromonte, Asif T. Chinwalla, Deanna M. Church, Michele Clamp, Christopher Clee, Francis S. Collins, Lisa L. Cook, Richard R. Copley, Alan Coulson, Olivier Couronne, James Cuff, Val Curwen, Tim Cutts, Mark Daly, Robert David, Joy Davies, Kimberly D. Delehaunty, Justin Deri, Emmanouil T. Dermitzakis, Colin Dewey, Nicholas J. Dickens, Mark Diekhans, Sheila Dodge, Inna Dubchak, Diane M. Dunn, Sean R. Eddy, Laura Elnitski, Richard D. Emes, Pallavi Eswara, Eduardo Eyras, Adam Felsenfeld, Ginger A. Fewell, Paul Flicek, Karen Foley, Wayne N. Frankel, Lucinda A. Fulton, Robert S. Fulton, Terrence S Furey, Diane Gage, Richard A. Gibbs, Gustavo Glusman, Sante Gnerre, Nick Goldman, Leo Goodstadt, Darren Grafham, Tina A. Graves, Eric D. Green, Simon Gregory, Roderic Guigó, Mark Guyer, Ross C. Hardison, David Haussler, Yoshihide Hayashizaki, LaDeana W. Hillier, Angela Hinrichs, Wratko Hlavina, Timothy Holzer, Fan Hsu, Axin Hua, Tim Hubbard, Adrienne Hunt, Ian Jackson, David B. Jaffe, L. Steven Johnson, Matthew Jones, Thomas A. Jones, Ann Joy, Michael Kamal, Elinor K. Karlsson, Donna Karolchik, Arkadiusz Kasprzyk, Jun Kawai, Evan Keibler, Cristyn Kells, W. James Kent, Andrew Kirby, Diana L. Kolbe, Ian Korf, Raju S. Kucherlapati, Edward J. Kulbokas, David Kulp, Tom Landers, J.P. Leger, Steven Leonard, Ivica Letunic, Rosie Levine, Jia Li, Ming Li, Christine Lloyd, Susan Lucas, Bin Ma, Donna R. Maglott, Elaine R. Mardis, Lucy Matthews, Evan Mauceli, John H Mayer, Megan McCarthy, W. Richard McCombie, Stuart McLaren, Kirsten McLay, John D. McPherson, Jim Meldrim, Beverley Meredith, Jill P Mesirov, Webb Miller, Tracie L. Miner, Emmanuel Mongin, Kate T. Montgomery, Michael Morgan, Richard Mott, James C. Mullikin, Donna M. Muzny, William E. Nash, Joanne O. Nelson, Michael N. Nhan, Robert Nicol, Zemin Ning, Chad Nusbaum, Michael J. O'Connor, Yasushi Okazaki, Karen Oliver, Emma Overton-Larty, Lior Pachter, Genís Parra, Kymberlie H. Pepin, Jane Peterson, Pavel Pevzner, Robert Plumb, Craig S. Pohl, Alex Poliakov, Tracy C. Ponce, Chris P. Ponting, Simon Potter, Michael Quail, Alexandre Reymond, Bruce A. Roe, Krishna M. Roskin, Edward M. Rubin, Alistair G. Rust, Ralph Santos, Victor Sapojnikov, Brian Schultz, Jörg Schultz, Matthias S. Schwartz, Scott Schwartz, Carol Scott, Steven Seaman, Steve Searle, Ted Sharpe, Andrew Sheridan, Ratna Shownkeen, Sarah Sims, Jonathan B. Singer, Guy Slater, Arian Smit, Douglas R. Smith, Brian Spencer, Arne Stabenau, Nicole Stange-Thomann, Charles Sugnet, Mikita Suyama, Glenn Tesler, Johanna Thompson, David Torrents, Evanne Trevaskis, John Tromp, Catherine Ucla, Abel Ureta-Vidal, Jade P. Vinson, Andrew C. Von Niederhausern, Claire M. Wade, Melanie Wall, Ryan J. Weber, Robert B. Weiss, Michael C. Wendl, Anthony P. West, Kris Wetterstrand, Raymond Wheeler, Simon Whelan, Jamey Wierzbowski, David Willey, Sophie Williams, Richard K. Wilson, Eitan Winter, Kim C. Worley, Dudley Wyman, Shan Yang, Shiaw-Pyng Yang, Evgeny M. Zdobnov, Michael C. Zody, and Eric S. Lander. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–562, 2002. ISSN 1476-4687. doi:

10.1038/nature01262.

[31] Juliana A. Ronchi, Tiago R. Figueira, Felipe G. Ravagnani, Helena C.F. Oliveira, Anibal E. Vercesi, and Roger F. Castilho. A spontaneous mutation in the nicotinamide nucleotide transhydrogenase gene of C57BL/6J mice results in mitochondrial redox abnormalities. *Free Radical Biology and Medicine*, 63:446–456, 2013. ISSN 1873-4596. doi: 10.1016/j.freeradbiomed.2013.05.049.

[32] Anne-Marie Aubin, Félix Lombard-Vadnais, Roxanne Collin, Holly A Aliesky, Sandra M McLachlan, and Sylvie Lesage. The NOD Mouse Beyond Autoimmune Diabetes. *Frontiers in Immunology*, 13:874769, 2022. ISSN 1664-3224. doi: 10.3389/fimmu.2022.874769.

[33] Hans-Georg Joost and Annette Schürmann. The genetic basis of obesity-associated type 2 diabetes (diabesity) in polygenic mouse models. *Mammalian Genome*, 25(9-10):401–412, 2014. ISSN 1432-1777. doi: 10.1007/s00335-014-9514-2.

[34] Steven A. Belinsky, Steven A. Stefanski, and Marshall W. Anderson. The A/J Mouse Lung as a Model for Developing New Chemointervention Strategies. *Cancer Research*, 53(2):410–416, 1993. ISSN 1538-7445.

[35] Vanessa De Vooght, Jeroen A. J. Vanoirbeek, Katrien Luyts, Steven Haenen, Benoit Nemery, and Peter H. M. Hoet. Choice of Mouse Strain Influences the Outcome in a Mouse Model of Chemical-Induced Asthma. *PLOS ONE*, 5(9):e12581, 2010. ISSN 1932-6203. doi: 10.1371/journal.pone.0012581.

[36] Tirumalai Rangasamy, Vikas Misra, Lijie Zhen, Clarke G. Tankersley, Rubin M. Tuder, and Shyam Biswal. Cigarette smoke-induced emphysema in A/J mice is associated with pulmonary oxidative stress, apoptosis of lung cells, and global alterations in gene expression. *The American Journal of Physiology - Lung Cellular and Molecular Physiology*, 296(6):L888–L900, 2009. ISSN 1522-1504. doi: 10.1152/ajplung.90369.2008.

[37] Mark A. Hornsey, Steven H. Laval, Rita Barresi, Hanns Lochmüller, and Kate Bushby. Muscular dystrophy in dysferlin-deficient mouse models. *Neuromuscular Disorders*, 23 (5):377–387, 2013. ISSN 0960-8966. doi: 10.1016/j.nmd.2013.02.004.

[38] Tiffany A. Garbutt, Thomas I. Konneker, Kranti Konganti, Andrew E. Hillhouse, Francis Swift-Haire, Alexis Jones, Drake Phelps, David L. Aylor, and David W. Threadgill. Permissiveness to form pluripotent stem cells may be an evolutionarily derived characteristic in Mus musculus. *Scientific Reports*, 8(1):14706, 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-32116-8.

[39] Ardian Ferraj, Peter A Audano, Parithi Balachandran, Anne Czechanski, Jacob I Flores, Alexander A Radecki, Varun Mosur, David S Gordon, Isha A Walawalkar, Evan E Eichler, Laura G Reinholdt, and Christine R. Beck. Resolution of structural variation in diverse mouse genomes reveals chromatin remodeling due to transposable elements. *Cell Genomics*, 3(5):100291, 2023. ISSN 2666-979X. doi: 10.1016/j.xgen.2023.100291.

[40] Peter M. Visscher, Naomi R. Wray, Qian Zhang, Pamela Sklar, Mark I. McCarthy, Matthew A. Brown, and Jian Yang. 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics*, 101(1):5–22, 2017. ISSN 1537-6605. doi: 10.1016/j.ajhg.2017.06.005.

[41] Luke J. O'Connor, Armin P. Schoech, Farhad Hormozdiari, Steven Gazal, Nick Pat-

terson, and Alkes L. Price. Extreme Polygenicity of Complex Traits Is Explained by Negative Selection. *American Journal of Human Genetics*, 105(3):456–476, 2019. ISSN 0002-9297. doi: 10.1016/j.ajhg.2019.07.003.

[42] The National Center for Biotechnology Information. Genome assembly grcm39, 2024. URL `https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001635.27/`.

[43] The National Center for Biotechnology Information. Mus musculus genomes, 2024. URL `https://api.ncbi.nlm.nih.gov/datasets/genome/?taxon=10090`.

[44] Megan Phifer-Rixey and Michael W. Nachman. Insights into mammalian biology from the wild house mouse Mus musculus. *eLife*, 15(4):e05959, 2015. ISSN 2050-084X. doi: 10.7554/eLife.05959.

[45] S. D. Ferris, R. D. Sage, E. M. Prager, U. Ritte, and A. C. Wilson. Mitochondrial DNA evolution in mice. *Genetics*, 105(3):681–721, 1983. ISSN 1943-2631. doi: 10.1093/genetics/105.3.681.

[46] Beth L. Dumont and Bret A. Payseur. Genetic analysis of genome-scale recombination rate evolution in house mice. *PLoS Genetics*, 7(6):e1002116, 2011. ISSN 1553-7404. doi: 10.1371/journal.pgen.1002116.

[47] Kelly A Frazer, Eleazar Eskin, Hyun Min Kang, Molly A Bogue, David A Hinds, Erica J Beilharz, Robert V Gupta, Julie Montgomery, Matt M Morenzoni, Geoffrey B Nilsen, Charit L Pethiyagoda, Laura L Stuve, Frank M Johnson, Mark J Daly, Claire M Wade, and David R Cox. A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. *Nature*, 448(7157):1050–1053, 2007. ISSN 0028-0836. doi: 10.1038/nature06067.

[48] Thomas M. Keane, Leo Goodstadt, Petr Danecek, Michael A. White, Kim Wong, Binnaz Yalcin, Andreas Heger, Avigail Agam, Guy Slater, Martin Goodson, Nicholas A. Furlotte, Eleazar Eskin, Christoffer Nellåker, Helen Whitley, James Cleak, Deborah Janowitz, Polinka Hernandez-Pliego, Andrew Edwards, T. Grant Belgard, Peter L. Oliver, Rebecca E. McIntyre, Amarjit Bhomra, Jérôme Nicod, Xiangchao Gan, Wei Yuan, Louise van der Weyden, Charles A. Steward, Sendu Bala, Jim Stalker, Richard Mott, Richard Durbin, Ian J. Jackson, Anne Czechanski, José Afonso Guerra-Assunção, Leah Rae Donahue, Laura G. Reinholdt, Bret A. Payseur, Chris P. Ponting, Ewan Birney, Jonathan Flint, and David J. Adams. Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature*, 477(7364):289–294, 2011. ISSN 1476-4687. doi: 10.1038/nature10413.

[49] Andrey A. Perelygin, Svetlana V. Scherbik, Igor B. Zhulin, Bronislava M. Stockman, Yan Li, and Margo A. Brinton. Positional cloning of the murine flavivirus resistance gene. *Proceedings of the National Academy of Sciences of the United States of America*, 99(14):9322–9327, 2002. ISSN 1091-6490. doi: 10.1073/pnas.142287799.

[50] Patricia L. Earl, Jeffrey L. Americo, and Bernard Moss. Insufficient Innate Immunity Contributes to the Susceptibility of the Castaneous Mouse to Orthopoxvirus Infection. *Journal of Virology*, 91(19):e01042–17, 2017. ISSN 1098-5514. doi: 10.1128/JVI.01042-17.

[51] Sarah R. Leist, Carolin Pilzner, Judith M. A. van den Brand, Leonie Dengler, Robert Geffers, Thijs Kuiken, Rudi Balling, Heike Kollmus, and Klaus Schughart. Influenza H3N2 infection of the collaborative cross founder strains reveals highly divergent host responses and identifies a unique phenotype in CAST/EiJ mice. *BMC Genomics*, 27

(17):143, 2016. ISSN 1471-2164. doi: 10.1186/s12864-016-2483-y.

[52] Takao Omura, Kumiko Omura, Andrea Tedeschi, Priscilla Riva, Michio W. Painter, Leticia Rojas, Joshua Martin, Véronique Lisi, Eric A. Huebner, Alban Latremoliere, Yuqin Yin, Lee B. Barrett, Bhagat Singh, Stella Lee, Tom Crisman, Fuying Gao, Songlin Li, Kush Kapur, Daniel H Geschwind, Kenneth S. Kosik, Giovanni Coppola, Zhigang He, S. Thomas Carmichael, Larry I. Benowitz, Michael Costigan, and Clifford J Woolf. Robust Axonal Regeneration Occurs in the Injured CAST/Ei Mouse CNS. *Neuron*, 86(5):1215–1227, 2015. ISSN 1097-4199. doi: 10.1016/j.neuron.2015.05.005.

[53] Laura E. Griffin, Lauren Essenmacher, Kathryn C. Racine, Lisard Iglesias-Carres, Jeffery S. Tessem, Susan M. Smith, and Andrew P. Neilson. Diet-induced obesity in genetically diverse collaborative cross mouse founder strains reveals diverse phenotype response and amelioration by quercetin treatment in 129S1/SvImJ, PWK/EiJ, CAST/PhJ, and WSB/EiJ mice. *The Journal of Nutritional Biochemistry*, (87): 108521, 2021. ISSN 0955-2863. doi: 10.1016/j.jnutbio.2020.108521.

[54] Fanny Odet, Wenqi Pan, Timothy A. Bell, Summer G. Goodson, Alicia M. Stevans, Zianing Yun, David L. Aylor, Chia-Yu Kao, Leonard McMillan, Fernando Pardo-Manuel de Villena, and Deborah A. O'Brien. The Founder Strains of the Collaborative Cross Express a Complex Combination of Advantageous and Deleterious Traits for Male Reproduction. *G3: Genes, Genomes, Genetics*, 5(12):2671–2683, 2015. ISSN 2160-1836. doi: 10.1534/g3.115.020172.

[55] James P. Noonan, Graham Coop, Sridhar Kudaravalli, Doug Smith, Johannes Krause, Joe Alessi, Feng Chen, Darren Platt, Svante Pääbo, Jonathan K. Pritchard, and Edward M. Rubin. Sequencing and Analysis of Neanderthal Genomic DNA. *Science*, 314 (5802):1113–1118, 2006. ISSN 1095-9203. doi: 10.1126/science.1131412.

[56] Cristina Sisu, Paul Muir, Adam Frankish, Ian Fiddes, Mark Diekhans, David Thybert, Duncan T. Odom, Paul Flicek, Thomas M. Keane, Tim Hubbard, Jennifer Harrow, and Mark Gerstein. Transcriptional activity and strain-specific history of mouse pseudogenes. *Nature Communications*, 11(1):3695, 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-17157-w.

[57] Erik Pettersson, Joakim Lundeberg, and Afshin Ahmadian. Generations of sequencing technologies. *Genomics*, 93(2):105–111, 2009. ISSN 0888-7543. doi: 10.1016/j.ygeno.2008.10.003.

[58] Gary A. Churchill, David C. Airey, Hooman Allayee, Joe M. Angel, Alan D. Attie, Jackson Beatty, William D. Beavis, John K. Belknap, Beth Bennett, Wade Berrettini, Andre Bleich, Molly Bogue, Karl W. Broman, Kari J. Buck, Ed Buckler, Margit Burmeister, Elissa J. Chesler, James M. Cheverud, Steven Clapcote, Melloni N. Cook, Roger D. Cox, John C. Crabbe, Wim E. Crusio, Ariel Darvasi, Christian F. Deschepper, R. W. Doerge, Charles R. Farber, Jiri Forejt, Daniel Gaile, Steven J. Garlow, Hartmut Geiger, Howard Gershenfeld, Terry Gordon, Jing Gu, Weikuan Gu, Gerald de Haan, Nancy L. Hayes, Craig Heller, Heinz Himmelbauer, Robert Hitzemann, Kent Hunter, Hui-Chen Hsu, Fuad A. Iraqi, Boris Ivandic, Howard J. Jacob, Ritsert C. Jansen, Karl J. Jepsen, Dabney K. Johnson, Thomas E Johnson, Gerd Kempermann, Christina Kendziorski, Malak Kotb, R. Frank Kooy, Bastien Llamas, Frank Lammert, Jean-Michel Lassalle, Pedro R Lowenstein, Lu Lu, Aldons Lusis, Ken-

neth F. Manly, Ralph Marcucio, Doug Matthews, Juan F. Medrano, Darla R. Miller, Guy Mittleman, Beverly A. Mock, Jeffrey S. Mogil, Xavier Montagutelli, Grant Morahan, David G. Morris, Richard Mott, Joseph H. Nadeau, Hiroki Nagase, Richard S. Nowakowski, Bruce F O'Hara, Alexander V. Osadchuk, Grier P. Page, Beverly Paigen, Kenneth Paigen, Abraham A. Palmer, Huei-Ju Pan, Leena Peltonen-Palotie, Jeremy Peirce, Daniel Pomp, Michal Pravenec, Daniel R. Prows, Zhonghua Qi, Roger H. Reeves, John Roder, Glenn D. Rosen, Eric E. Schadt, Leonard C. Schalkwyk, Ze'ev Seltzer, Kazuhiro Shimomura, Siming Shou, Mikko J. Sillanpää, Linda D. Siracusa, Hans-Willem Snoeck, Jimmy L. Spearow, Karen Svenson, Lisa M. Tarantino, David Threadgill, Linda A. Toth, William Valdar, Fernando Pardo-Manuel de Villena, Craig Warden, Steve Whatley, Robert W. Williams, Tim Wiltshire, Nengjun Yi, Dabao Zhang, Min Zhang, and Fei Zou. The Collaborative Cross, a community resource for the genetic analysis of complex traits. *Nature Genetics*, 36(36):1133–1137, 2004. ISSN 1546-1718. doi: 10.1038/ng1104-1133.

[59] David L. Aylor, William Valdar, Wendy Foulds-Mathes, Ryan J. Buus, Ricardo A. Verdugo, Ralph S. Baric, Martin T. Ferris, Jeff A. Frelinger, Mark Heise, Matt B. Frieman, Lisa E. Gralinski, Timothy A. Bell, John D. Didion, Kunjie Hua, Derrick L. Nehrenberg, Christine L. Powell, Jill Steigerwalt, Yuying Xie, Samir N.P. Kelada, Francis S. Collins, Ivana V. Yang, David A. Schwartz, Lisa A. Branstetter, Elissa J. Chesler, Darla R. Miller, Jason Spence, Eric Yi Liu, Leonard McMillan, Abhishek Sarkar, Jeremy Wang, Wei Wang, Qi Zhang, Karl W. Broman, Ron Korstanje, Caroline Durrant, Richard Mott, Fuad A. Iraqi, Daniel Pomp, David Threadgill, Fernando Pardo-Manuel de Villena, and Gary A. Churchill. Genetic analysis of complex traits in the emerging Collaborative Cross. *Genome Research*, 21(8):1213–1222, 2011. ISSN 1549-5469. doi: 10.1101/gr.111310.110.

[60] Adam Roberts, Fernando Pardo-Manuel de Villena, Wei Wang, Leonard McMillan, and David W. Threadgill. The polymorphism architecture of mouse genetic resources elucidated using genome-wide resequencing data: implications for QTL discovery and systems genetics. *Mammalian Genome*, 18(6-7):473–481, 2007. ISSN 1432-1777. doi: 10.1007/s00335-007-9045-1.

[61] David W Threadgill, Darla R Miller, Gary A Churchill, and Fernando Pardo-Manuel de Villena. The collaborative cross: a recombinant inbred mouse population for the systems genetic era. *ILAR Journal*, 51(1):24–31, 2011. ISSN 1930-6180. doi: 10.1093/ilar.52.1.24.

[62] Karen L. Svenson, Daniel M. Gatti, William Valdar, Catherine E. Welsh, Riyan Cheng, Elissa J. Chesler, Abraham A. Palmer, Leonard McMillan, and Gary A. Churchill. High-resolution genetic mapping using the Mouse Diversity outbred population. *Genetics*, 190(2):437–447, 2012. ISSN 1943-2631. doi: 10.1534/genetics.111.132597.

[63] David W Threadgill, Kent W Hunter, and Robert W Williams. Genetic dissection of complex and quantitative traits: from fantasy to reality via a community effort. *Mammalian Genome*, 13(4):175–178, 2002. ISSN 0938-8990. doi: 10.1007/s00335-001-4001-Y.

[64] Fuad A. Iraqi, Mustafa Mahajne, Yasser Salaymah, Hani Sandovski, Hanna Tayem, Karin Vered, Lois Balmer, Michael Hall, Glynn Manship, Grant Morahan, Ken Pettit, Jeremy Scholten, Kathryn Tweedie, Andrew Wallace, Lakshini Weerasekera, James

Cleak, Caroline Durrant, Leo Goodstadt, Richard Mott, Binnaz Yalcin, David L. Aylor, Ralph S. Baric, Timothy A. Bell, Katharine M. Bendt, Jennifer Brennan, Jackie D. Brooks, Ryan J. Buus, James J. Crowley, John D. Calaway, Mark E. Calaway, Agnieszka Cholka, David B. Darr, John P. Didion, Amy Dorman, Eric T. Everett, Martin T. Ferris, Wendy Foulds Mathes, Chen-Ping Fu, Terry J. Gooch, Summer G. Goodson, Lisa E. Gralinski, Stephanie D. Hansen, Mark T. Heise, Jane Hoel, Kunjie Hua, Mayanga C. Kapita, Seunggeun Lee, Alan B. Lenarcic, Eric Yi Liu, Hedi Liu, Leonard McMillan, Terry R. Magnuson, Kenneth F. Manly, Darla R. Miller, Deborah A. O'Brien, Fanny Odet, Isa Kemal Pakatci, Wenqi Pan, Fernando Pardo-Manuel de Villena, Charles M. Perou, Daniel Pomp, Corey R. Quackenbush, Nashiya N. Robinson, Norman E. Sharpless, Ginger D. Shaw, Jason S. Spence, Patrick F. Sullivan, Wei Sun, Lisa M. Tarantino, William Valdar, Jeremy Wang, Wei Wang, Catherine E. Welsh, Alan Whitmore, Tim Wiltshire, Fred A. Wright, Yuying Xie, Zaining Yun, Vasyl Zhabotynsky, Zhaojun Zhang, Fei Zou, Christine Powell, Jill Steigerwalt, David W. Threadgill, Elissa J. Chesler, Gary A. Churchill, Daniel M. Gatti, Ron Korstanje, Karen L. Svenson, Francis S. Collins, Nigel Crawford, Kent Hunter, Samir N. P. Kelada, Bailey C. E. Peck, Karlyne Reilly, Urraca Tavarez, Daniel Bottomly, Robert Hitzeman, Shannon K. McWeeney, Jeffrey Frelinger, Harsha Krovi, Jason Phillippi, Richard A. Spritz, Lauri Aicher, Michael Katze, Elizabeth Rosenzweig, Ariel Shusterman, Aysar Nashef, Ervin I. Weiss, Yael Houri-Haddad, Morris Solle, Robert W. Williams, Klaus Schughart, Hyuna Yang, John E. French, Andrew K. Benson, Jaehyoung Kim, Ryan Legge, Soo Jen Low, Fangrui Ma, Ines Martinez, Jens Walter, Karl W. Broman, Benedikt Hallgrimsson, Ophir Klein, George Weinstock, Wesley C. Warren, Yvana V. Yang, and David. Schwartz. The Genome Architecture of the Collaborative Cross Mouse Genetic Reference Population. *Genetics*, 190(2):389–401, 2012. ISSN 1943-2631. doi: 10.1534/genetics.111.132639.

[65] Gregory R. Keele, Bryan C. Quach, Jennifer W. Israel, Grace A. Chappell, Lauren Lewis, Alexias Safi, Jeremy M. Simon, Paul Cotney, Gregory E. Crawford, William Valdar, Ivan Rusyn, and Terrence S. Furey. Integrative QTL analysis of gene expression and chromatin accessibility identifies multi-tissue patterns of genetic regulation. *PLoS Genetics*, 16(1):e1008537, 2020. ISSN 1553-7404. doi: 10.1371/journal.pgen.1008537.

[66] Gary A. Churchill, Daniel M. Gatti, Steven C. Munger, and Karen L. Svenson. The Diversity Outbred mouse population. *Mammalian Genome*, 23(9-10):713–718, 2012. ISSN 1432-1777. doi: 10.1007/s00335-012-9414-2.

[67] Basel M. Al-Barghouthi, Larry D. Mesner, Gina M. Calabrese, Daniel Brooks, Steven M. Tommasini, Mary L. Bouxsein, Mark C. Horowitz, Clifford J. Rosen, Kevin Nguyen, Samuel Haddox, Emily A. Farber, Suna Onengut-Gumuscu, Daniel Pomp, and Charles R. Farber. Systems genetics in diversity outbred mice inform BMD GWAS and identify determinants of bone strength. *Nature Communications*, 12(1):3408, 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-23649-0.

[68] Luke J. Dillard, Will T. Rosenow, Gina M. Calabrese, Larry D. Mesner, Basel M. Al-Barghouthi, Abdullah Abood, Emily A. Farber, Suna Onengut-Gumuscu, Steven M. Tommasini, Mark A. Horowitz, Clifford J. Rosen, Lutian Yao, Ling Qin, and Charles R. Farber. Single-Cell Transcriptomics of Bone Marrow Stromal Cells in Diversity Outbred Mice: A Model for Population-Level scRNA-Seq Studies. *Journal of Bone and*

Mineral Research, 38(9):1350–1363, 2023. ISSN 1523-4681. doi: 10.1002/jbmr.4882.

[69] Kristen J. Nowak and Kay E. Davies. Duchenne muscular dystrophy and dystrophin: pathogenesis and opportunities for treatment. *EMBO reports*, 5(9):872–876, 2004. ISSN 1469-3178. doi: 10.1038/sj.embor.7400221.

[70] Amandine Barral and Jérôme Déjardin. The chromatin signatures of enhancers and their dynamic regulation. *Nucleus*, 14(1):2160551, 2023. ISSN 1949-1042. doi: 10.1080/19491034.2022.2160551.

[71] Zhijian Li, Marcel H. Schulz, Thomas Look, Matthias Begemann, Martin Zenke, and Ivan G. Costa. Identification of transcription factor binding sites using ATAC-seq. *Genome Biology*, 20(1):45, 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1642-2.

[72] Jeanette Reinartz, Eddy Bruyns, Jing-Zhong Lin, Tim Burcham, Sydney Brenner, Ben Bowen, Michael Kramer, and Rick Woychik. Massively parallel signature sequencing (MPSS) as a tool for in-depth quantitative gene expression profiling in all organisms. *Briefings in Functional Genomics and Proteomics*, 1(1):95–104, 2002. ISSN 1473-9550. doi: 10.1093/bfgp/1.1.95.

[73] Mark Schena. Genome analysis with gene expression microarrays. *BioEssays*, 18(5):427–431, 1996. ISSN 1521-1878. doi: 10.1002/bies.950180513.

[74] Jingyue Ju, Dae Hyun Kim, Lanrong Bi, Qinglin Meng, Xiaopeng Bai, Zengmin Li, Xiaoxu Li, Mong Sano Marma, Shundi Shi, Jian Wu, John R. Edwards, Aireen Romu, and Nicholas J. Turro. Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *Proceedings of the National Academy of Sciences of the United States of America*, 103(52):19635–19640, 2006. ISSN 1091-6490. doi: 10.1073/pnas.0609513103.

[75] Taishan Hu, Nilesh Chitnis, Dimitri Monos, and Anh Dinh. Next-generation sequencing technologies: An overview. *Human Immunology*, 82(11):801–811, 2021. ISSN 0198-8859. doi: 10.1016/j.humimm.2021.02.012.

[76] Ali Mortazavi, Brian A. Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):621–628, 2008. ISSN 1548-7091. doi: 10.1038/nmeth.1226.

[77] Simone Picelli, Omid R. Faridani, Asa K. Björklund, Gösta Winberg, Sven Sagasser, and Rickard Sandberg. Full-length RNA-seq from single cells using Smart-seq2. *Nature Protocols*, 9(1):171–181, 2014. ISSN 1754-2189. doi: 10.1038/nprot.2014.006.

[78] Yunhao Wang, Yue Zhao, Audrey Bollas, Yuru Wang, and Kin Fai Au. Nanopore sequencing technology, bioinformatics and applications. *Nature Biotechnology*, 39(11):1348–1365, 2021. ISSN 1087-0156. doi: 10.1038/s41587-021-01108-x.

[79] Lingyun Song and Gregory E. Crawford. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor Protocols*, 2:pdb.prot5384., 2010. ISSN 1559-6095. doi: 10.1101/pdb.prot5384.

[80] Fiorella C. Grandi, Hailey Modi, Lucas Kampman, and M. Ryan Corces. Chromatin accessibility profiling by ATAC-seq. *Nature Protocols*, 17(6):1518–1552, 2022. ISSN 1750-2799. doi: 10.1038/s41596-022-00692-9.

[81] ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, 306(5696):636–640, 2004. ISSN 1095-9203. doi: 10.1126/science.1105136.

[82] Matthew T. Maurano, Richard Humbert, Eric Rynes, Robert E. Thurman, Eric Haugen, Hao Wang, Alex P. Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody, Anthony Shafer, Fidencio Neri, Kristen Lee, Tanya Kutyavin, Sandra Stehling-Sun, Audra K. Johnson, Theresa K. Canfield, Erika Giste, Morgan Diegel, Daniel Bates, R. Scott Hansen, Shane Neph, Peter J. Sabo, Shelly Heimfeld, Antony Raubitschek, Steven Ziegler, Chris Cotsapas, Nona Sotoodehnia, Ian Glass, Shamil R. Sunyaev, Rajinder Kaul, and John A. Stamatoyannopoulos. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science*, 337(6099):1190–1195, 2012. ISSN 1095-9203. doi: 10.1126/science.1222794.

[83] Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J. Ziller, Viren Amin, John W. Whitaker, Matthew D. Schultz, Lucas D. Ward, Abhishek Sarkar, Gerald Quon, Richard S. Sandstrom, Matthew L. Eaton, Yi-Chieh Wu, Andreas R. Pfenning, Xinchen Wang, Melina Claussnitzer, Yaping Liu, Cristian Coarfa, R. Alan Harris, Noam Shoresh, Charles B. Epstein, Elizabeta Gjoneska, Danny Leung, Wei Xie, R. David Hawkins, Ryan Lister, Chibo Hong, Philippe Gascard, Andrew J. Mungall, Richard Moore, Eric Chuah, Angela Tam, Theresa K. Canfield, R. Scott Hansen, Rajinder Kaul, Peter J. Sabo, Mukul S. Bansal, Annaick Carles, Jesse R. Dixon, Kai-How Farh, Soheil Feizi, Rosa Karlic, Ah-Ram Kim, Ashwinikumar Kulkarni, Daofeng Li, Rebecca Lowdon, GiNell Elliott, Tim R. Mercer, Shane J. Neph, Vitor Onuchic, Paz Polak, Nisha Rajagopal, Pradipta Ray, Richard C. Sallari, Kyle T. Siebenthall, Nicholas A. Sinnott-Armstrong, Michael Stevens, Robert E. Thurman, Jie Wu, Bo Zhang, Xin Zhou, Arthur E. Beaudet, Laurie A. Boyer, Philip L. De Jager, Peggy J. Farnham, Susan J. Fisher, David Haussler, Steven J. M. Jones, Wei Li, Marco A. Marra, Michael T. McManus, Shamil Sunyaev, James A. Thomson, Thea D. Tlsty, Li-Huei Tsai, Wei Wang, Robert A. Waterland, Michael Q. Zhang, Lisa H. Chadwick, Bradley E. Bernstein, Joseph F. Costello, Joseph R. Ecker, Martin Hirst, Alexander Meissner, Aleksandar Milosavljevic, Bing Ren, John A. Stamatoyannopoulos, Ting Wang, and Manolis Kellis. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, 2015. ISSN 1476-4687. doi: 10.1038/nature14248.

[84] Peng He, Brian A. Williams, Diane Trout, Georgi K. Marinov, Henry Amrhein, Libera Berghella, Say-Tar Goh, Ingrid Plajzer-Frick, Veena Afzal, Len A. Pennacchio, Diane E. Dickel, Axel Visel, Bing Ren, Ross C. Hardison, Yu Zhang, and Barbara J. Wold. The changing mouse embryo transcriptome at whole tissue and single-cell resolution. *Nature*, 583(7818):760–767, 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2536-x.

[85] Wolfgang J. Streit. Microglia and Alzheimer's disease pathogenesis. *Journal of Neuroscience Research*, 77(1):1–8, 2004. ISSN 1097-4547. doi: 10.1002/jnr.20093.

[86] Byungjin Hwang, Ji Hyun Lee, and Duhee Bang. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental Molecular Medicine*, 50(8):1–14, 2018. ISSN 1226-3613. doi: 10.1038/s12276-018-0071-8.

[87] 10x Genomics. 10x genomics homepage, 2024. URL `https://www.10xgenomics.com/`.

[88] Alexander B. Rosenberg, Charles M. Roco, Richard A. Muscat, Anna Kuchina, Paul Sample, Zizhen Yao, Lucas T. Graybuck, David J. Peeler, Sumit Mukherjee, Wei Chen, Suzie H. Pun, Drew L. Sellers, Bosiljka Tasic, and Georg Seelig. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science*, 360

(6385):176–182, 2018. ISSN 1095-9203. doi: 10.1126/science.aam8999.

[89] Parse Biosciences. Parse biosciences homepage, 2024. URL `https://www.parsebiosciences.com/`.

[90] Weihua Zeng, Shan Jiang, Xiangduo Kong, Nicole El-Ali, Alexander R. Ball Jr., Christopher I-Hsing Ma, Naohiro Hashimoto, Kyoko Yokomori, and Ali Mortazavi. Single-nucleus RNA-seq of differentiating human myoblasts reveals the extent of fate heterogeneity. *Nucleic Acids Research*, 44(21):e158, 2016. ISSN 1362-4962. doi: 10.1093/nar/gkw739.

[91] Sai Ma, Bing Zhang, Lindsay M. LaFave, Andrew S. Earl, Zachary Chiang, Yan Hu, Jiarui Ding, Alison Brack, Vinay K. Kartha, Tristan Tay, Travis Law, Caleb Lareau, Ya-Chieh Hsu, Aviv Regev, and Jason D. Buenrostro. Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell*, 184(3):1103–1116.e20, 2020. ISSN 1097-4172. doi: 10.1016/j.cell.2020.09.056.

[92] Jeongwoo Lee, Do Young Hyeon, and Daehee Hwang. Single-cell multiomics: technologies and data analysis methods. *Experimental Molecular Medicine*, 52(9):1428–1442, 2020. ISSN 1226-3613. doi: 10.1038/s12276-020-0420-2.

[93] Teemu Kivioja, Anna Vähärautio, Kasper Karlsson, Martin Bonke, Martin Enge, Sten Linnarsson, and Jussi Taipale. Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods*, 9(1):72–74, 2011. ISSN 1548-7091. doi: 10.1038/nmeth.1778.

[94] Benjamin Kaminow, Dinar Kaminow, and Alexander Dobin. STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. *bioRxiv*, page 2021.05.05.442755, 2021. doi: 10.1101/2021.05.05.442755.

[95] Páll Melsted, A Sina Booeshaghi, Lauren Liu, Fan Gao, Lambda Lu, Kyung Hoi Joseph Min, Eduardo da Veiga Beltrame, Kristján Eldjárn Hjörleifsson, Jase Gehring, and Lior Pachter. Modular, efficient and constant-memory single-cell RNA-seq preprocessing. *Nature Biotechnology*, 39(7):813–818, 2021. ISSN 1087-0156. doi: 10.1038/s41587-021-00870-2.

[96] Nayoung Kim, Huiram Kang, Areum Jo, Seung-Ah Yoo, and Hae-Ock Lee. Perspectives on single-nucleus RNA sequencing in different cell types and tissues. *Journal of Pathology and Translational Medicine*, 57(1):52–59, 2023. ISSN 2383-7845. doi: 10.4132/jptm.2022.12.19.

[97] Lopez Romain Wolock, Samuel L. and Allon M. Klein. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Systems*, 8(4):281–291, 2019. ISSN 2405-4720. doi: 10.1016/j.cels.2018.11.005.

[98] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M Mauck 3rd, Shiwei Zheng, Andrew Butler, Maddie J Lee, Aaron J Wilk, Charlotte Darby, Michael Zager, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P Mimitou, Jaison Jain, Avi Srivastava, Tim Stuart, Lamar M Fleming, Bertrand Yeung, Angela J Rogers, Juliana M McElrath, Catherine A Blish, Raphael Gottardo, Peter Smibert, and Rahul Satija. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587, 2021. ISSN 1097-4172. doi: 10.1016/j.cell.2021.04.048.

[99] F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15, 2018. ISSN 1474-760X. doi: 10.1186/s13059-017-1382-0.

[100] V. A. Traag, L. Waltman, and N. J. van Eck. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1):5233, 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-41695-z.

[101] L. McInnes, J. Healy, N. Saul, and L. Großberger. UMAP: Uniform Manifold Approximation and Projection. *J Open Source Software*, 2018. ISSN 2475-9066. doi: 10.21105/joss.00861.

[102] Tara Chari and Lior Pachter. The specious art of single-cell genomics. *PLOS Computational Biology*, 19(8):e1011288, 2023. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1011288.

[103] Changde Cheng, Wenan Chen, Hongjian Jin, and Xiang Chen. A Review of Single-Cell RNA-Seq Annotation, Integration, and Cell-Cell Communication. *Cells*, 12(15):1970, 2023. ISSN 2073-4409. doi: 10.3390/cells12151970.

[104] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M 3rd Mauck, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive Integration of Single-Cell Data. *Cell*, 177(7):1888–1902, 2019. ISSN 1097-4172. doi: 10.1016/j.cell.2019.05.031.

[105] Jeffrey T. Wigle, Natasha Harvey, Michael Detmar, Irina Lagutina, Gerard Grosveld, Michael D. Gunn, David G. Jackson, and Guillermo Oliver. An essential role for Prox1 in the induction of the lymphatic endothelial cell phenotype. *The EMBO Journal*, 21 (7):1505–1513, 2002. ISSN 1460-2075. doi: 10.1093/emboj/21.7.1505.

[106] Alfonso Lavado, Oleg V. Lagutin, Lionel M. L. Chow, Suzanne J. Baker, and Guillermo Oliver. Prox1 Is Required for Granule Cell Maturation and Intermediate Progenitor Maintenance During Brain Neurogenesis. *PLOS Biology*, 8(8):e1000460, 2010. ISSN 1545-7885. doi: 10.1371/journal.pbio.1000460.

[107] Chen-Che Jeff Huang and Yuan Kang. The transient cortical zone in the adrenal gland: the mystery of the adrenal X-zone. *Journal of Endocrinology*, 241(1):R51–R63, 2019. ISSN 1687-8337. doi: 10.1530/JOE-18-0632.

[108] Tijana Radic, Lara Frieß, Aruvi Vijikumar, Tassilo Jungenitz, Thomas Deller, and Stephan W. Schwarzacher. Differential Postnatal Expression of Neuronal Maturation Markers in the Dentate Gyrus of Mice and Rats. *Frontiers in Neuroanatomy*, 11(104), 2017. ISSN 1662-5129. doi: 10.3389/fnana.2017.00104.

[109] Hongkui Zeng. What is a cell type and how to define it? *Cell*, 185(15):2739–2755, 2023. ISSN 1097-4172. doi: 10.1016/j.cell.2022.06.031.

[110] Jonas Simon Fleck, J. Gray Camp, and Barbara Treutlein. What is a cell type? *Science*, 381(6659):733–734, 2023. ISSN 1095-9203. doi: 10.1126/science.adf6162.

[111] Hugo C. Olguin, Zhihong Yang, Stephen J Tapscott, and Bradley B. Olwin. Reciprocal inhibition between Pax7 and muscle regulatory factors modulates myogenic cell fate determination. *Journal of Cell Biology*, 177(5):769–779, 2007. ISSN 0021-9525. doi: 10.1083/jcb.200608122.

[112] Michael J. Petrany, Casey O. Swoboda, Chengyi Sun, Kashish Chetal, Xiaoting Chen, Matthew T. Weirauch, Nathan Salomonis, and Douglas P. Millay. Single-nucleus RNA-seq identifies transcriptional heterogeneity in multinucleated skeletal myofibers. *Nature Communications*, 11(1):6374, 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-20063-w.

[113] Matthieu Dos Santos, Stéphanie Backer, Benjamin Saintpierre, Brigitte Izac, Muriel

Andrieu, Franck Letourneur, Frederic Relaix, Athanassia Sotiropoulos, and Pascal Maire. Single-nucleus RNA-seq and FISH identify coordinated transcriptional activity in mammalian myofibers. *Nature Communications*, 11(1):5102, 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-18789-8.

[114] Elisa Negroni, Maria Kondili, Laura Muraine, Mona Bensalah, Gillian Sandra Butler-Browne, Vincent Mouly, Anne Bigot, and Capucine Trollet. Muscle fibro-adipogenic progenitors from a single-cell perspective: Focus on their "virtual" secretome. *Frontiers in Cell and Developmental Biology*, 10(952041), 2022. ISSN 2296-634X. doi: 10.3389/fcell.2022.952041.

[115] Feng Yan, David R. Powell, David J. Curtis, and Nicholas C. Wong. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biology*, 21(1):22, 2020. ISSN 1474-760X. doi: 10.1186/s13059-020-1929-3.

[116] Rashida Rajgara, Hamood AlSudais, Aisha Saleh, Alex Brown, Ines Barrakad, Alexandre Blais, and Nadine Wiper-Bergeron. The glucocorticoid receptor is a critical regulator of muscle satellite cell quiescence. *bioRxiv*, page 2023.08.27.555012, 2023. doi: 10.1101/2023.08.27.555012.

[117] Bruno Cadot, Vincent Gache, and Edgar R. Gomes. Moving and positioning the nucleus in skeletal muscle - one step at a time. *Nucleus*, 6(5):373–381, 2015. ISSN 1949-1042. doi: 10.1080/19491034.2015.1090073.

[118] Katherine Williams, Kyoko Yokomori, and Ali Mortazavi. Heterogeneous Skeletal Muscle Cell and Nucleus Populations Identified by Single-Cell and Single-Nucleus Resolution Transcriptome Assays. *Frontiers in Genetics*, 2022. ISSN 1664-8021. doi: 10.3389/fgene.2022.835099.

[119] Matthieu Dos Santos, Stéphanie Backer, Benjamin Saintpierre, Brigitte Izac, Muriel Andrieu, Franck Letourneur, Frederic Relaix, Athanassia Sotiropoulos, and Pascal Maire. Single-nucleus RNA-seq and FISH identify coordinated transcriptional activity in mammalian myofibers. *Nature Communications*, 11(1):5102, 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-18789-8.

[120] John B. Scott, Catherine L. Ward, Benjamin T. Corona, Michael R. Deschenes, Benjamin S. Harrison, Justin M. Saul, and George J. Christ. Achieving Acetylcholine Receptor Clustering in Tissue-Engineered Skeletal Muscle Constructs In vitro through a Materials-Directed Agrin Delivery Approach. *Frontiers in Pharmacology*, 7(508), 2017. ISSN 1663-9812. doi: 10.3389/fphar.2016.00508.

[121] Daniel L. Plotkin, Michael D. Roberts, Cody T. Haun, and Brad J. Schoenfeld. Muscle Fiber Type Transitions with Exercise Training: Shifting Perspectives. *Sports*, 9(9):127, 2021. ISSN 2075-4663. doi: 10.3390/sports9090127.

[122] Peter Langfelder and Steve Horvath. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 29(9):559, 2008. ISSN 1471-2105. doi: 10.1186/1471-2105-9-559.

[123] Samuel Morabito, Fairlie Reese, Negin Rahimzadeh, Emily Miyoshi, and Vivek Swarup. hdWGCNA identifies co-expression networks in high-dimensional transcriptomics data. *Cell Reports Methods*, 3(6):100498, 2023. ISSN 2667-2375. doi: 10.1016/j.crmeth.2023.100498.

[124] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003. doi: 10.5555/944919.944937.

[125] Jonathan K. Pritchard, Matthew Stephens, and Peter Donnelly. Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, 155(2):945–959, 2000. ISSN 1943-2631. doi: 10.1093/genetics/155.2.945.

[126] Xiaotian Wu, Hao Wu, and Zhijin Wu. Penalized Latent Dirichlet Allocation Model in Single-Cell RNA Sequencing. *Statistics in Biosciences*, 13:543–562, 2021. ISSN 1867-1772. doi: 10.1007/s12561-021-09304-8.

[127] Qi Yang, Zhaochun Xu, Wenyang Zhou, Pingping Wang, Qinghua Jiang, and Liran Juan. An interpretable single-cell RNA sequencing data clustering method based on latent Dirichlet allocation. *Briefings in Bioinformatics*, 24(4):bbad199, 2023. ISSN 1477-4054. doi: 10.1093/bib/bbad199.

[128] Narges Rezaie, Elisabeth Rebboah, Brian A. Williams, Heidi Yahan Liang, Fairlie Reese, Gabriela Balderrama-Gutierrez, Louise A. Dionne, Laura Reinholdt, Diane Trout, Barbara J. Wold, and Ali Mortazavi. Identification of robust cellular programs using reproducible Latent Dirichlet Allocation. *bioRxiv*, page 2024.02.26.582178, 2023. doi: 10.1101/2024.02.26.582178.

[129] Camille Bouchard and Jacques P Tremblay. Portrait of Dysferlinopathy: Diagnosis and Development of Therapy. *Journal of Clinical Medicine*, 12(18):6011, 2023. ISSN 2077-0383. doi: 10.3390/jcm12186011.

[130] Roger Volden, Theron Palmer, Ashley Byrne, Charles Cole, Robert J. Schmitz, Richard E. Green, and Christopher Vollmers. Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA. *Proceedings of the National Academy of Sciences*, 115(39):9726–9731, 2018. ISSN 1091-6490. doi: 10.1073/pnas.1806447115.

[131] Ishaan Gupta, Paul G. Collier, Bettina Haase, Ahmed Mahfouz, Anoushka Joglekar, Taylor Floyd, Frank Koopmans, Ben Barres, August B. Smit, Steven A. Sloan, Wenjie Luo, Olivier Fedrigo, M. Elizabeth Ross, and Hagen U. Tilgner. Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nature Biotechnology*, 36:1197–1202, 2018. ISSN 1546-1696. doi: 10.1038/nbt.4259.

[132] Xiaoying Fan, Dong Tang, Yuhan Liao, Pidong Li, Yu Zhang, Minxia Wang, Fan Liang, Xiao Wang, Yun Gao, Lu Wen, Depeng Wang, Yang Wang, and Fuchou Tang. Single-cell RNA-seq analysis of mouse preimplantation embryos by third-generation sequencing. *PLOS Biology*, 18(12):e3001017, 2020. ISSN 1545-7885. doi: 10.1371/journal.pbio.3001017.

[133] Anoushka Joglekar, Andrey Prjibelski, Ahmed Mahfouz, Paul Collier, Susan Lin, Anna Katharina Schlusche, Jordan Marrocco, Stephen R. Williams, Bettina Haase, Ashley Hayes, Jennifer G. Chew, Neil I. Weisenfeld, Man Ying Wong, Alexander N. Stein, Simon A. Hardwick, Toby Hunt, Qi Wang, Christoph Dieterich, Zachary Bent, Olivier Fedrigo, Steven A. Sloan, Davide Risso, Erich D. Jarvis, Paul Flicek, Wenjie Luo, Geoffrey S. Pitt, Adam Frankish, August B. Smit, M. Elizabeth Ross, and Hagen U. Tilgner. A spatially resolved brain region- and cell type-specific isoform atlas of the postnatal mouse brain. *Nature Communications*, 12(1):463, 2021. ISSN 2041-1723. doi: 10.1038/s41467-020-20343-5.

[134] Z. Wang and P. J. Grabowski. Cell- and stage-specific splicing events resolved in specialized neurons of the rat cerebellum. *RNA*, 2(12):1241–1253, 1996. ISSN 1469-9001.

[135] Martin A. Smith and Diane K. O'Dowd. Cell-specific regulation of agrin RNA splicing in the chick ciliary ganglion. *Neuron*, 12(4):795–804, 1994. ISSN 1097-4199. doi: 10.1016/0896-6273(94)90332-8.

[136] S. Lawrence Zipursky, Woj M. Wojtowicz, and Daisuke Hattori. Got diversity? Wiring the fly brain with Dscam. *Trends in Biochemical Sciences*, 31(10):581–588, 2006. ISSN 0968-0004. doi: 10.1016/j.tibs.2006.08.003.

[137] Marcus Frank and Rolf Kemler. Protocadherins. *Current Opinion in Cell Biology*, 14 (5):557–562, 2002. ISSN 0955-0674. doi: 10.1016/s0955-0674(02)00365-4.

[138] Tony Kwan, David Benovoy, Christel Dias, Scott Gurd, Cathy Provencher, Patrick Beaulieu, Thomas J Hudson, Rob Sladek, and Jacek Majewski. Genome-wide analysis of transcript isoform variation in humans. *Nature Genetics*, 40(2):225–231, 2008. ISSN 1546-1718. doi: 10.1038/ng.2007.57.

[139] Robert J Osborne and Charles A. Thornton. RNA-dominant diseases. *Human Molecular Genetics*, 15(2):R162–R169, 2006. ISSN 1460-2083. doi: 10.1093/hmg/ddl181.

[140] A. V. Philips, L. T. Timchenko, and T. A. Cooper. Disruption of splicing regulated by a CUG-binding protein in myotonic dystrophy. *Science*, 280(5364):737–741, 1998. ISSN 1095-9203. doi: 10.1126/science.280.5364.737.

[141] Xiaoyan Lin, Jill W. Miller, Ami Mankodi, Rahul N. Kanadia, Yuan Yuan, Richard T. Moxley, Maurice S. Swanson, and Charles A. Thornton. Failure of MBNL1-dependent post-natal splicing transitions in myotonic dystrophy. *Human Molecular Genetics*, 15 (13):2087–2097, 2006. ISSN 1460-2083. doi: 10.1093/hmg/ddl132.

[142] D.D.W. Cornelison. Context Matters: In Vivo and In Vitro Influences on Muscle Satellite Cell Activity. *Journal of Cellular Biochemistry*, 105(3):663–669, 2013. ISSN 0730-2312. doi: 10.1002/jcb.21892.

[143] Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J. Lennon, Kenneth J. Livak, Tarjei S. Mikkelsen, and John L. Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, 32(4):381–386, 2014. ISSN 1546-1696. doi: 10.1038/nbt.2859.

[144] Norio Motohashi and Atsushi Asakura. Muscle satellite cell heterogeneity and self-renewal. *Frontiers in Cell and Developmental Biology*, 2(1), 2014. ISSN 2296-634X. doi: 10.3389/fcell.2014.00001.

[145] Cole Trapnell, Brian A. Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J. van Baren, Steven L. Salzberg, Barbara J. Wold, and Lior Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5):511–515, 2010. ISSN 1546-1696. doi: 10.1038/nbt.1621.

[146] Y Harada, M Nakamura, and A. Asano. Temporally distinctive changes of alternative splicing patterns during myogenic differentiation of C2C12 cells. *The Journal of Biochemistry*, 118(4):780–790, 1995. ISSN 1520-4995. doi: 10.1093/oxfordjournals.jbchem.a124980.

[147] Balderrama-Gutierrez G. Reese F. Jiang S. Rahmanian S. Zeng W. Williams B. Trout D. England W. Chu S. Spitale R. C. Tenner A. Wold B. Mortazavi A. Wyman, D. A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification. *bioRxiv*, page 672931, 2019. doi: 10.1101/672931.

[148] Manuel Tardaguila, Lorena de la Fuente, Cristina Marti, Cécile Pereira, Francisco Jose Pardo-Palacios, Hector Del Risco, Marc Ferrell, Maravillas Mellado, Marissa Macchietto, Kenneth Verheggen, Mariola Edelmann, Iakes Ezkurdia, Jesus Vazquez, Michael Tress, Ali Mortazavi, Lennart Martens, Susana Rodriguez-Navarro, Victoria Moreno-Manzano, and Ana Conesa. SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Research*, 28(3):396–411, 2018. ISSN 1549-5469. doi: 10.1101/gr.222976.117.

[149] Gioele La Manno, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, Maria E Kastriti, Peter Lönnerberg, Alessandro Furlan, Jean Fan, Lars E Borm, Zehua Liu, David van Bruggen, Jimin Guo, Xiaoling He, Roger Barker, Erik Sundström, Gonçalo Castelo-Branco, Patrick Cramer, Igor Adameyko, Sten Linnarsson, and Peter V. Kharchenko. RNA velocity of single cells. *Nature*, 560(7719):494–498, 2018. ISSN 1476-4687. doi: 10.1038/s41586-018-0414-6.

[150] Fairlie Reese and Ali Mortazavi. Swan: a library for the analysis and visualization of long-read transcriptomes. *Bioinformatics*, 37(9):1322–1323, 2021. ISSN 1460-2059. doi: 10.1093/bioinformatics/btaa836.

[151] Irina Neganova, Felipe Vilella, Stuart P Atkinson, Maria Lloret, João F Passos, Thomas von Zglinicki, José-Enrique O'Connor, Deborah Burks, Richard Jones, Lyle Armstrong, and Majlinda Lako. An important role for CDK2 in G1 to S checkpoint activation and DNA damage response in human embryonic stem cells. *Stem Cells*, 29 (4):651–659, 2011. ISSN 1066-5099. doi: 10.1002/stem.620.

[152] W. Strzalka and A. Ziemienowicz. Proliferating cell nuclear antigen (PCNA): a key factor in DNA replication and cell cycle regulation. *Annals of Botany*, 107(7):1127–1140, 2011. ISSN 0305-7364. doi: 10.1093/aob/mcq243.

[153] T Chen, Y Sun, P Ji, S Kopetz, and W. Zhang. Topoisomerase II in chromosome instability and personalized cancer therapy. *Oncogene*, 34(31):4019–4031, 2015. ISSN 1476-5594. doi: 10.1038/onc.2014.332.

[154] Hongxia Ren, Ping Yin, and Cunming Duan. IGFBP-5 regulates muscle cell differentiation by binding to IGF-II and switching on the IGF-II auto-regulation loop. *Journal of Cell Biology*, 182(5):979–991, 2008. ISSN 0021-9525. doi: 10.1083/jcb.200712110.

[155] Giuliana Rossi, Stefania Antonini, Chiara Bonfanti, Stefania Monteverde, Chiara Vezzali, Shahragim Tajbakhsh, Giulio Cossu, and Graziella Messina. Nfix Regulates Temporal Progression of Muscle Regeneration through Modulation of Myostatin Expression. *Cell Reports*, 14(9):2238–2249, 2016. ISSN 2211-1247. doi: 10.1016/j.celrep.2016.02.014.

[156] C Florian Bentzinger, Yu Xin Wang, Julia von Maltzahn, Vahab D Soleimani, Hang Yin, and Michael A. Rudnicki. Fibronectin regulates Wnt7a signaling and satellite cell expansion. *Cell Stem Cell*, 12(1):75–87, 2013. ISSN 1934-5909. doi: 10.1016/j.stem.2012.09.015.

[157] Kelsey Thomas, Adam J Engler, and Gretchen A. Meyer. Extracellular matrix regulation in the muscle satellite cell niche. *Connective Tissue Research*, 56(1):1–8, 2015. ISSN 0300-8207. doi: 10.3109/03008207.2014.947369.

[158] Moon-Chang Choi, Soyoung Ryu, Rui Hao, Bin Wang, Meghan Kapur, Chen-Ming Fan, and Tso-Pang Yao. HDAC4 promotes Pax7-dependent satellite cell activation

and muscle regeneration. *EMBO Reports*, 15(11):1175–1183, 2014. ISSN 1469-3178. doi: 10.15252/embr.201439195.

[159] Stephanie N. Oprescu, Feng Yue, Jiamin Qiu, Luiz F. Brito, and Shihuan Kuang. Temporal Dynamics and Heterogeneity of Cell Populations during Skeletal Muscle Regeneration. *iScience*, 23(4):100993, 2020. ISSN 2589-0042. doi: 10.1016/j.isci.2020.100993.

[160] Tadashi Yoshida. MCAT elements and the TEF-1 family of transcription factors in muscle development and disease. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 28(1):8–17, 2008. ISSN 1524-4636. doi: 10.1161/ATVBAHA.107.155788.

[161] Tim Stuart, Avi Srivastava, Shaista Madad, Caleb A Lareau, and Rahul Satija. Single-cell chromatin state analysis with Signac. *Nature Methods*, 18(11):1333–1341, 2021. ISSN 1548-7105. doi: 10.1038/s41592-021-01282-5.

[162] Bin Wei and J-P Jin. TNNT1, TNNT2, and TNNT3: Isoform genes, regulation, and structure-function relationships. *Gene*, 582(1):1–13, 2016. ISSN 1879-0038. doi: 10.1016/j.gene.2016.01.006.

[163] Matthew S. Hestand, Andreas Klingenhoff, Matthias Scherf, Yavuz Ariyurek, Yolande Ramos, Wilbert van Workum, Makoto Suzuki, Thomas Werner, Gert-Jan B. van Ommen, Johan T. den Dunnen, Matthias Harbers, and Peter A. C. 't Hoen. Tissue-specific transcript annotation and expression profiling with complementary next-generation sequencing technologies. *Nucleic Acids Research*, 38(16):e165, 2010. ISSN 1362-4962. doi: 10.1093/nar/gkq602.

[164] ENCODE Project Consortium, Jill E. Moore, Michael J. Purcaro, Henry E. Pratt, Charles B. Epstein, Noam Shoresh, Jessika Adrian, Trupti Kawli, Carrie A. Davis, Alexander Dobin, Rajinder Kaul, Jessica Halow, Eric L. Van Nostrand, Peter Freese, David U. Gorkin, Yin Shen, Yupeng He, Mark Mackiewicz, Florencia Pauli-Behn, Brian A. Williams, Ali Mortazavi, Cheryl A. Keller, Xiao-Ou Zhang, Shaimae I. Elhajjajy, Jack Huey, Diane E. Dickel, Valentina Snetkova, Xintao Wei, Xiaofeng Wang, Juan Carlos Rivera-Mulia, Joel Rozowsky, Jing Zhang, Surya B. Chhetri, Jialing Zhang, Alec Victorsen, Kevin P. White, Axel Visel, Gene W. Yeo, Christopher B. Burge, Eric Lécuyer, David M. Gilbert, Job Dekker, John Rinn, Eric M. Mendenhall, Joseph R. Ecker, Manolis Kellis, Robert J. Klein, William S. Noble, Anshul Kundaje, Roderic Guigó, Peggy J. Farnham, J. Michael Cherry, Richard M. Myers, Bing Ren, Brenton R. Graveley, Mark B. Gerstein, Len A. Pennacchio, Michael P. Snyder, Bradley E. Bernstein, Barbara Wold, Ross C. Hardison, Thomas R. Gingeras, John A. Stamatoyannopoulos, and Zhiping. Weng. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, 583(7818):699–710, 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2493-4.

[165] Riikka Kivelä, Ida Salmela, Yen Hoang Nguyen, Tatiana V Petrova, Heikki A Koistinen, Zoltan Wiener, and Kari Alitalo. The transcription factor Prox1 is essential for satellite cell differentiation and muscle fibre-type regulation. *Nature Communications*, 12(7):13124, 2016. ISSN 2041-1723. doi: 10.1038/ncomms13124.

[166] Riikka Kivelä, Ida Salmela, Yen Hoang Nguyen, Tatiana V Petrova, Heikki A Koistinen, Zoltan Wiener, and Kari Alitalo. Dual function of VGLL4 in muscle regeneration. *The EMBO Journal*, 38(17):e101051, 2019. ISSN 1460-2075. doi: 10.15252/embj.2018101051.

[167] Bio-Rad Laboratories Inc. Illumina Bio-Rad SureCell® WTA 3' Library Prep Kit for Nuclei Samples. Document Number 1000000044178 Ver. 00. 2018. doi: https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/surecell/surecell-wta3-nuclei-demonstrated-protocol-1000000044178-00.pdf.

[168] Bio-Rad Laboratories Inc. Illumina Bio-Rad SureCell® ATAC-Seq Library Preparation Kit User Guide. Document Number 100000106678 Ver. 1.0.1. 2018. doi: https://www.bio-rad.com/webroot/web/pdf/lsr/literature/10000106678.pdf.

[169] Advanced Cell Diagnostics Inc. Cultured adherent cells sample preparation for RNAscope® Multiplex Fluorescent v2. Document Number MK-50-010. 2019. doi: https://acdbio.com/technical-support/user-manuals.

[170] Advanced Cell Diagnostics Inc. RNAscope® HiPlex assay with sample preparation and pretreatment. Document Number 324100-USM. 2019. doi: https://acdbio.com/technical-support/user-manuals.

[171] Advanced Cell Diagnostics Inc. Tech Note for using RNAscope® HiPlex Alternative Display Module. Document Number MK-51-132. 2019. doi: https://acdbio.com/technical-support/user-manuals.

[172] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34 (18):3094–3100, 2018. ISSN 1367-4811. doi: 10.1093/bioinformatics/bty191.

[173] Dana Wyman and Ali Mortazavi. TranscriptClean: variant-aware correction of indels, mismatches and splice junctions in long-read transcripts. *Bioinformatics*, 35(2):340–342, 2019. ISSN 1367-4811. doi: 10.1093/bioinformatics/bty483.

[174] Fairlie Reese and Roger Volden. fairliereese/LR-splitpipe: LR-splitpipe v1.0 (v1.0). *Zenodo*, 2021. doi: 10.5281/zenodo.5168059.

[175] Fairlie Reese. fairliereese/2021_c2c12: c2c12 figure code (v1.0). *Zenodo*, 2021. doi: 10.5281/zenodo.5168057.

[176] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 2013. ISSN 1367-4811. doi: 10.1093/bioinformatics/bts635.

[177] Rahul Satija, Jeffrey A. Farrell, David Gennert, Alexander F. Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, 33(5): 495–502, 2015. ISSN 1546-1696. doi: 10.1038/nbt.3192.

[178] Christoph Hafemeister and Rahul Satija. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology*, 20(1):296, 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1874-1.

[179] Bio-Rad Laboratories Inc. Bio-Rad ATAC-Seq Analysis Toolkit Tutorial. Document Number 7191 Ver. 1.0.0. 2019. doi: https://www.bio-rad.com/webroot/web/pdf/lsr/literature/Bulletin_7191.pdf.

[180] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp324.

[181] Yong Zhang, Tao Liu, Clifford A. Meyer, Jérôme Eeckhoute, David S. Johnson, Bradley E. Bernstein, Chad Nusbaum, Richard M. Myers, Myles Brown, Wei Li, and X. Shirley Liu. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology*, 9(9):

R137, 2008. ISSN 1474-760X. doi: 10.1186/gb-2008-9-9-r137.

[182] Wenbao Yu, Yasin Uzun, Qin Zhu, Changya Chen, and Kai Tan. scATAC-pro: a comprehensive workbench for single-cell chromatin accessibility sequencing data. *Genome Biology*, 21(1):94, 2020. ISSN 1474-760X. doi: 10.1186/s13059-020-02008-0.

[183] Fidel Ramírez, Friederike Dündar, Sarah Diehl, Björn A. Grüning, and Thomas Manke. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Research*, 42:W187–W191, 2014. ISSN 1362-4962. doi: 10.1093/nar/gku365.

[184] Cory Y. McLean, Dave Bristor, Michael Hiller, Shoa L. Clarke, Bruce T. Schaar, Craig B. Lowe, Aaron M. Wenger, and Gill Bejerano. GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology*, 28(5):495–501, 2010. ISSN 1546-1696. doi: 10.1038/nbt.1630.

[185] Edward Y. Chen, Christopher M. Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gabriela Vaz Meirelles, Neil R. Clark, and Avi Ma'ayan. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, 14 (128), 2013. ISSN 1471-2105. doi: 10.1186/1471-2105-14-128.

[186] Aaron R. Quinlan and Ira M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010. ISSN 1367-4811. doi: 10.1093/bioinformatics/btq033.

[187] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet journal*, 17(1), 2011. ISSN 2226-6089. doi: 10.14806/ej.17.1.200.

[188] Hestand MS and 't Hoen PA. DeepCAGE and DeepSAGE with proliferating and differentiated C2C12 mouse myoblasts. *GEO*, 2010. doi: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE21580.

[189] Shane Neph, M. Scott Kuehn, Alex P. Reynolds, Eric Haugen, Robert E. Thurman, Audra K. Johnson, Eric Rynes, Matthew T. Maurano, Jeff Vierstra, Sean Thomas, Richard Sandstrom, Richard Humbert, and John A. Stamatoyannopoulos. BEDOPS: high-performance genomic feature operations. *Bioinformatics*, 28(14):1919–1920, 2012. ISSN 1367-4811. doi: 10.1093/bioinformatics/bts277.

[190] A. S. Hinrichs, D. Karolchik, R. Baertsch, G. P. Barber, G. Bejerano, H. Clawson, M. Diekhans, T. S. Furey, R. A. Harte, F. Hsu, J. Hillman-Jackson, R. M. Kuhn, J. S. Pedersen, A. Pohl, B. J. Raney, K. R. Rosenbloom, A. Siepel, K. E. Smith, C. W. Sugnet, A. Sultan-Qurraie, D. J. Thomas, H. Trumbower, R. J. Weber, M. Weirauch, A. S. Zweig, D. Haussler, and W. J. Kent. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Research*, 1(34):D590–D598, 2006. ISSN 1362-4962. doi: 10.1093/nar/gkj144.

[191] Wencheng Li, Bei You, Mainul Hoque, Dinghai Zheng, Wenting Luo, Zhe Ji, Ji Yeon Park, Samuel I. Gunderson, Auinash Kalsotra, James L Manley, and Bin Tian. Systematic profiling of poly(A)+ transcripts modulated by core 3' end processing and splicing factors reveals regulatory rules of alternative cleavage and polyadenylation. *PLOS Genetics*, 11(4):e1005166, 2015. ISSN 1553-7404. doi: 10.1371/journal.pgen.1005166.

[192] Wencheng Li, Bei You, Dinghai Zheng, Mainul Hoque, Wenting Luo, Zhe Ji, Ji Yeon Park, and Bin Tian. Regulation of alternative cleavage and polyadenylation by 3' end processing and splicing factors. *GEO*, 2015. doi: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE62001.

[193] Adam Frankish, Mark Diekhans, Anne-Maud Ferreira, Rory Johnson, Irwin Jungreis,

Jane Loveland, Jonathan M Mudge, Cristina Sisu, James Wright, Joel Armstrong, If Barnes, Andrew Berry, Alexandra Bignell, Silvia Carbonell Sala, Jacqueline Chrast, Fiona Cunningham, Tomás Di Domenico, Sarah Donaldson, Ian T. Fiddes, Carlos García Girón, Jose Manuel Gonzalez, Tiago Grego, Matthew Hardy, Thibaut Hourlier, Toby Hunt, Osagie G. Izuogu, Julien Lagarde, Fergal J. Martin, Laura Martínez, Shamika Mohanan, Paul Muir, Fabio C. P. Navarro, Anne Parker, Baikang Pei, Fernando Pozo, Magali Ruffier, Bianca M. Schmitt, Eloise Stapleton, Marie-Marthe Suner, Irina Sycheva, Barbara Uszczynska-Ratajczak, Jinuri Xu, Andrew Yates, Daniel Zerbino, Yan Zhang, Bronwen Aken, Jyoti S. Choudhary, Mark Gerstein, Roderic Guigó, Tim J. P Hubbard, Manolis Kellis, Benedict Paten, Alexandre Reymond, Michael L. Tress, and Paul Flicek. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research*, 47(D1):D766–D773, 2019. ISSN 1362-4962. doi: 10.1093/nar/gky955.

[194] Tabula Muris Consortium. A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature*, 583(7817):590–595, 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2496-1.

[195] Tabula Muris Consortium. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature*, 562(7727):367–372, 2018. ISSN 1476-4687. doi: 10.1038/s41586-018-0590-4.

[196] Zizhen Yao, Cindy T. J. van Velthoven, Thuc Nghi Nguyen, Jeff Goldy, Adriana E. Sedeno-Cortes, Fahimeh Baftizadeh, Darren Bertagnolli, Tamara Casper, Megan Chiang, Kirsten Crichton, Song-Lin Ding, Olivia Fong, Emma Garren, Alexandra Glandon, Nathan W. Gouwens, James Gray, Lucas T. Graybuck, Michael J. Hawrylycz, Daniel Hirschstein, Matthew Kroll, Kanan Lathia, Changkyu Lee, Boaz Levi, Delissa McMillen, Stephanie Mok, Thanh Pham, Qingzhong Ren, Christine Rimorin, Nadiya Shapovalova, Josef Sulc, Susan M. Sunkin, Michael Tieu, Amy Torkelson, Herman Tung, Katelyn Ward, Nick Dee, Kimberly A. Smith, Bosiljka Tasic, and Hongkui Zeng. A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. *Cell*, 184(12):3222–3241, 2021. ISSN 1097-4172. doi: 10.1016/j.cell.2021.04.021.

[197] Narges Rezaie, Fairlie Reese, and Ali Mortazavi. PyWGCNA: a Python package for weighted gene co-expression network analysis. *Bioinformatics*, 39(7):btad415, 2023. ISSN 1460-2059. doi: 10.1093/bioinformatics/btad415.

[198] Kazumasa Kanemaru, James Cranley, Daniele Muraro, Antonio M. A. Miranda, Siew Yen Ho, Anna Wilbrey-Clark, Jan Patrick Pett, Krzysztof Polanski, Laura Richardson, Monika Litvinukova, Natsuhiko Kumasaka, Yue Qin, Zuzanna Jablonska, Claudia I. Semprich, Lukas Mach, Monika Dabrowska, Nathan Richoz, Liam Bolt, Lira Mamanova, Rakeshlal Kapuge, Sam N. Barnett, Shani Perera, Carlos Talavera-López, Ilaria Mulas, Krishnaa T. Mahbubani, Liz Tuck, Lu Wang, Margaret M. Huang, Martin Prete, Sophie Pritchard, John Dark, Kourosh Saeb-Parsy, Minal Patel, Menna R. Clatworthy, Norbert Hübner, Rasheda A. Chowdhury, Michela Noseda, and Sarah A. Teichmann. Spatially resolved multiomics of human cardiac niches. *Nature*, 619(7971): 801–810, 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06311-1.

[199] Micheal A. McLellan, Daniel A. Skelly, Malathi S.I. Dona, Galen T. Squiers, Gabriella E. Farrugia, Taylah L. Gaynor, Charles D. Cohen, Raghav Pandey, Henry Diep, Antony Vinh, Nadia A. Rosenthal, and Alexander R. Pinto. High-Resolution

Transcriptomic Profiling of the Heart During Chronic Stress Reveals Cellular Drivers of Cardiac Fibrosis and Hypertrophy. *Circulation*, 142(15):1448–1463, 2020. ISSN 1524-4539. doi: 10.1161/CIRCULATIONAHA.119.045115.

[200] Michael J. Petrany, Casey O. Swoboda, Chengyi Sun, Kashish Chetal, Xiaoting Chen, Matthew T. Weirauch, Nathan Salomonis, and Douglas P. Millay. Single-nucleus RNA-seq identifies transcriptional heterogeneity in multinucleated skeletal myofibers. *Nature Communications*, 11(1):6374, 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-20063-w.

[201] Hitoshi Ishimoto and Robert B. Jaffe. Development and Function of the Human Fetal Adrenal Cortex: A Key Component in the Feto-Placental Unit. *Endocrine Reviews*, 32(3):317–355, 2011. ISSN 1945-7189. doi: 10.1210/er.2010-0001.

[202] W. M. van Weerden, H. G. Bierings, G. J. van Steenbrugge, F. H. de Jong, and F. H. Schröder. Adrenal glands of mouse and rat do not synthesize androgens. *Life Sciences*, 50(12):857–861, 1992. ISSN 1879-0631. doi: 10.1016/0024-3205(92)90204-3.

[203] Gerd Kempermann, Hongjun Song, and Fred H. Gage. Neurogenesis in the Adult Hippocampus. *Cold Spring Harbor Perspectives in Biology*, 7(9):a018812, 2015. ISSN 1943-0264. doi: 10.1101/cshperspect.a018812.

[204] Peter S. Eriksson, Ekaterina Perfilieva, Thomas Björk-Eriksson, Ann-Marie Alborn, Claes Nordborg, Daniel A. Peterson, and Fred H Gage. Neurogenesis in the adult human hippocampus. *Nature Medicine*, 4(11):1313–1317, 1998. ISSN 1546-170X. doi: 10.1038/3305.

[205] Gabriel Berdugo-Vega, Shonali Dhingra, and Federico Calegari. Sharpening the blades of the dentate gyrus: how adult-born neurons differentially modulate diverse aspects of hippocampal learning and memory. *The EMBO Journal*, page e113524, 2023. doi: 10.15252/embj.2023113524.

[206] Eri Segi-Nishida and Kanzo Suzuki. Regulation of adult-born and mature neurons in stress response and antidepressant action in the dentate gyrus of the hippocampus. *Neuroscience Research*, S0168-0102(22):00233–00234, 2022. doi: 10.1016/j.neures.2022.08.010.

[207] Jeong Beom Kim, Hyunah Lee, Marcos J Araúzo-Bravo, Kyujin Hwang, Donggyu Nam, Myung Rae Park, Holm Zaehres, Kook In Park, and Seok-Jin Lee. Oct4-induced oligodendrocyte progenitor cells enhance functional recovery in spinal cord injury model. *The EMBO Journal*, 34(23):2971–2983, 2015. ISSN 1460-2075. doi: 10.15252/embj.201592652.

[208] Leslie Kirby, Jing Jin, Jaime Gonzalez Cardona, Matthew D. Smith, Kyle A. Martin, Jingya Wang, Hayley Strasburger, Leyla Herbst, Maya Alexis, Jodi Karnell, Todd Davidson, Ranjan Dutta, Joan Goverman, Dwight Bergles, and Peter A. Calabresi. Oligodendrocyte precursor cells present antigen and are cytotoxic targets in inflammatory demyelination. *Nature Communications*, 10(1):3887, 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-11638-3.

[209] Martin Leu, Elisabeth Ehler, and J.-C. Perriard. Characterisation of postnatal growth of the murine heart. *Anatomy and embryology*, 204(3):217–214, 2001. ISSN 0340-2061. doi: 10.1007/s004290100206.

[210] Zhenlong Xin, Zhiqiang Ma, Shuai Jiang, Dongjin Wang, Chongxi Fan, Shouyin Di, Wei Hu, Tian Li, Junjun She, and Yang Yang. FOXOs in the impaired heart: New

therapeutic targets for cardiac diseases. *Biochimica et Biophysica Acta*, 1863(2):486–498, 2017. ISSN 0006-3002. doi: 10.1016/j.bbadis.2016.11.023.

[211] Carsten Skurk, Yasuhiro Izumiya, Henrike Maatz, Peter Razeghi, Ichiro Shiojima, Marco Sandri, Kaori Sato, Ling Zeng, Stephan Schiekofer, David Pimentel, Stewart Lecker, Heinrich Taegtmeyer, Alfred L. Goldberg, and Kenneth Walsh. The FOXO3a transcription factor regulates cardiac myocyte size downstream of AKT signaling. *Journal of Biological Chemistry*, 280(21):20814–20823, 2005. ISSN 1083-351X. doi: 10.1074/jbc.M500528200.

[212] Alexandra Wiesinger, Gerard J. J. Boink, Vincent M. Christoffels, and Harsha D. Devalla. Retinoic acid signaling in heart development: Application in the differentiation of cardiovascular lineages from human pluripotent stem cells. *Stem Cell Reports*, 16 (11):2589–2606, 2021. ISSN 2213-6711. doi: 10.1016/j.stemcr.2021.09.010.

[213] Toru Oka, Marjorie Maillet, Alistair J. Watt, Robert J. Schwartz, Bruce J. Aronow, Stephen A. Duncan, and Jeffery D. Molkentin. Cardiac-specific deletion of Gata4 reveals its requirement for hypertrophy, compensation, and myocyte viability. *Circulation Research*, 98(6):837–845, 2006. ISSN 1524-4571. doi: 10.1161/01.RES.0000215985.18538.c4.

[214] Egbert Bisping, Sadakatsu Ikeda, Sek Won Kong, Oleg Tarnavski, Natalya Bodyak, Julie R McMullen, Satish Rajagopal, Jennifer K Son, Qing Ma, Zhangli Springer, Peter M Kang, Seigo Izumo, and William T. Pu. Gata4 is required for maintenance of postnatal cardiac function and protection from pressure overload-induced heart failure. *PNAS*, 103(39):14471–14476, 2006. ISSN 0027-8424. doi: 10.1073/pnas.0602543103.

[215] Cody A Desjardins and Francisco J. Naya. The Function of the MEF2 Family of Transcription Factors in Cardiac Development, Cardiogenomics, and Direct Reprogramming. *Journal of Cardiovascular Development and Disease*, 3(3):26, 2016. ISSN 2308-3425. doi: 10.3390/jcdd3030026.

[216] Amira Moustafa, Sara Hashemi, Gurnoor Brar, Jörg Grigull, Siemon H. S. Ng, Declan Williams, Gerold Schmitt-Ulms, and John C. McDermott. The MEF2A transcription factor interactome in cardiomyocytes. *Cell Death Disease*, 14(4):240, 2023. ISSN 2041-4889. doi: 10.1038/s41419-023-05665-8.

[217] Stephanie L Padula, Nivedhitha Velayutham, and Katherine E. Yutzey. Transcriptional Regulation of Postnatal Cardiomyocyte Maturation and Regeneration. *International Journal of Molecular Sciences*, 22(6):3288, 2021. ISSN 1422-0067. doi: 10.3390/ijms22063288.

[218] Hugo C. Olguin, Zhihong Yang, Stephen J. Tapscott, and Bradley B Olwin. Reciprocal inhibition between Pax7 and muscle regulatory factors modulates myogenic cell fate determination. *Journal of Cell Biology*, 177(5):769–779, 2007. ISSN 0021-9525. doi: 0.1083/jcb.200608122.

[219] Alexis R. Demonbreun, Bridget H. Biersmith, and Elizabeth M. McNally. Membrane fusion in muscle development and repair. *Seminars in Cell Developmental Biology*, 45:48–56, 2015. ISSN 1096-3634. doi: 10.1016/j.semcdb.2015.10.026.

[220] Stefano Schiaffino, Alberto C Rossi, Vika Smerdu, Leslie A Leinwand, and Carlo Reggiani. Developmental myosins: expression patterns and functional significance. *Skeletal Muscle*, 5(22), 2015. ISSN 2044-5040. doi: 10.1186/s13395-015-0046-6.

[221] J Sher and C. Cardasis. Skeletal muscle fiber types in the adult mouse. *Acta Neurologica*

Scandinavica, 54(1):45–56, 1976. doi: 10.1111/j.1600-0404.1976.tb07619.x.

[222] Guy Karlebach and Ron Shamir. Modelling and analysis of gene regulatory networks. *Nature Reviews Molecular Cell Biology*, 9(10):770–780, 2008. ISSN 1471-0080. doi: 10.1038/nrm2503.

[223] Sina A. Booeshaghi, Ingileif B. Hallgrímsdóttir, Ángel Gálvez-Merchán, and Lior Pachter. Depth normalization for single-cell genomics count data. *bioRxiv*, page 2022.05.06.490859, 2022. doi: 10.1101/2022.05.06.490859.

[224] Kyle L MacQuarrie, Abraham P Fong, Randall H Morse, and Stephen J. Tapscott. Genome-wide transcription factor binding: beyond direct target regulation. *Trend in Genetics*, 27(4):141–148, 2011. ISSN 1362-4555. doi: 10.1016/j.tig.2011.01.001.

[225] Jason Gertz, Daniel Savic, Katherine E. Varley, E. Christopher Partridge, Alexias Safi, Preti Jain, Gregory M. Cooper, Timothy E. Reddy, Gregory E. Crawford, and Richard M. Myers. Distinct properties of cell type-specific and shared transcription factor binding sites. *Molecular Cell*, 52(1):25–36, 2014. ISSN 1097-2765. doi: 10.1016/j.molcel.2013.08.037.

[226] Nicole Ludwig, Petra Leidinger, Kurt Becker, Christina Backes, Tobias Fehlmann, Christian Pallasch, Steffi Rheinheimer, Benjamin Meder, Cord Stähler, Eckart Meese, and Andreas Keller. Distribution of miRNA expression across human tissues. *Nucleic Acids Research*, 44(8):3865–3877, 2016. ISSN 1362-4962. doi: 10.1093/nar/gkw116.

[227] Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-Ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*, 16(12):1289–1296, 2019. ISSN 1548-7091. doi: 10.1038/s41592-019-0619-0.

[228] Yinyi Wu and Karen K. Hirschi. Tissue-Resident Macrophage Development and Function. *Frontiers in Cell and Developmental Biology*, 8(8):617879, 2021. ISSN 2296-634X. doi: 10.3389/fcell.2020.617879.

[229] Chris S. Vink, Samanta A. Mariani, and Elaine Dzierzak. Embryonic Origins of the Hematopoietic System: Hierarchies and Heterogeneity. *Hemasphere*, 6(6):e737, 2022. ISSN 2572-9241. doi: 10.1097/HS9.0000000000000737.

[230] Yuki Hattori. The behavior and functions of embryonic microglia. *Anatomical Science International*, 97(1):1–14, 2022. ISSN 1447-073X. doi: 10.1007/s12565-021-00631-w.

[231] Bryce A Durafourt, Craig S Moore, Domenick A Zammit, Trina A Johnson, Fatma Zaguia, Marie-Christine Guiot, Amit Bar-Or, and Jack P. Antel. Comparison of polarization properties of human adult microglia and blood-derived macrophages. *Glia*, 60(5):717–727, 2012. ISSN 1098-1136. doi: 10.1002/glia.22298.

[232] K. Williams, A. Bar-Or, E. Ulvestad, A. Olivier, J. P. Antel, and V. W. Yong. Biology of adult human microglia in culture: comparisons with peripheral blood monocytes and astrocytes. *Journal of Neuropathology Experimental Neurology*, 51(5):538–549, 1992. doi: 10.1097/00005072-199209000-00009.

[233] Naoki Abe, Mohammed E. Choudhury, Minori Watanabe, Shun Kawasaki, Tasuku Nishihara, Hajime Yano, Shirabe Matsumoto, Takehiro Kunieda, Yoshiaki Kumon, Toshihiro Yorozuya, and Junya Tanaka. Comparison of the detrimental features of microglia and infiltrated macrophages in traumatic brain injury: A study using a hypnotic bromovalerylurea. *Glia*, 66(10):2158–2173, 2018. ISSN 1098-1136. doi: 10.1002/glia.23469.

[234] Roman Günthner and Hans-Joachim Anders. Interferon-regulatory factors determine macrophage phenotype polarization. *Mediators of Inflammation*, 2013:731023, 2013. ISSN 1466-1861. doi: 10.1155/2013/731023.

[235] Zhiyuan Vera Zheng, Junfan Chen, Hao Lyu, Sin Yu Erica Lam, Gang Lu, Wai Yee Chan, and George K C. Wong. Novel role of STAT3 in microglia-dependent neuroinflammation after experimental subarachnoid haemorrhage. *Stroke and Vascular Neurology*, 7(1):62–70, 2022. ISSN 2694-5746. doi: 10.1136/svn-2021-001028.

[236] Guoqiang Zhang, Jianan Lu, Jingwei Zheng, Shuhao Mei, Huaming Li, Xiaotao Zhang, An Ping, Shiqi Gao, Yuanjian Fang, and Jun Yu. Spi1 regulates the microglial/macrophage inflammatory response via the PI3K/AKT/mTOR signaling pathway after intracerebral hemorrhage. *Neural Regeneration Research*, 19(1):161–170, 2024. ISSN 1876-7958. doi: 10.4103/1673-5374.375343.

[237] Xi Xiao, Yuanyuan Hou, Wei Yu, and Sihua Qi. Propofol Ameliorates Microglia Activation by Targeting MicroRNA-221/222-IRF2 Axis. *Journal of Immunology Research*, page 3101146, 2021. ISSN 2314-7156. doi: 10.1155/2021/3101146.

[238] Bethany R. Fixsen, Claudia Z. Han, Yi Zhou, Nathanael J. Spann, Payam Saisan, Zeyang Shen, Christopher Balak, Mashito Sakai, Isidoro Cobo, Inge R. Holtman, Anna S. Warden, Gabriela Ramirez, Jana G. Collier, Martina P. Pasillas, Miao Yu, Rong Hu, Bin Li, Sarah Belhocine, David Gosselin, Nicole G. Coufal, Bing Ren, and Christopher K. Glass. SALL1 enforces microglia-specific DNA binding and function of SMADs to establish microglia identity. *Nature Immunology*, 24(7):1188–1199, 2023. ISSN 1529-2916. doi: 10.1038/s41590-023-01528-8.

[239] Sebastian G Utz, Peter See, Wiebke Mildenberger, Morgane Sonia Thion, Aymeric Silvin, Mirjam Lutz, Florian Ingelfinger, Nirmala Arul Rayan, Iva Lelios, Anne Buttgereit, Kenichi Asano, Shyam Prabhakar, Sonia Garel, Burkhard Becher, Florent Ginhoux, and Melanie Greter. Early Fate Defines Microglia and Non-parenchymal Brain Macrophage Development. *Cell*, 181(3):557–573, 2020. ISSN 1097-4172. doi: 10.1016/j.cell.2020.03.021.

[240] Edsel M Abud, Ricardo N Ramirez, Eric S Martinez, Luke M Healy, Cecilia H H Nguyen, Sean A Newman, Andriy V Yeromin, Vanessa M Scarfone, Samuel E Marsh, Cristhian Fimbres, Chad A Caraway, Gianna M Fote, Abdullah M Madany, Anshu Agrawal, Rakez Kayed, Karen H Gylys, Michael D Cahalan, Brian J Cummings, Jack P Antel, Ali Mortazavi, Monica J Carson, Wayne W Poon, and Mathew Blurton-Jones. iPSC-Derived Human Microglia-like Cells to Study Neurological Diseases. *Neuron*, 94 (2):278–293, 2017. ISSN 1097-4199. doi: 10.1016/j.neuron.2017.03.042.

[241] Elham Poonaki, Ulf Dietrich Kahlert, Sven G Meuth, and Ali Gorji. The role of the ZEB1-neuroinflammation axis in CNS disorders. *Journal of Neuroinflammation*, 19 (1):275, 2022. ISSN 1742-2094. doi: 10.1186/s12974-022-02636-2.

[242] Ana Estecha, Noemí Aguilera-Montilla, Paloma Sánchez-Mateos, and Amaya Puig-Kröger. RUNX3 regulates intercellular adhesion molecule 3 (ICAM-3) expression during macrophage differentiation and monocyte extravasation. *PLOS One*, 7(3):e33313, 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0033313.

[243] Kai Yan, Tian-Tian Da, Zhen-Hua Bian, Yi He, Meng-Chu Liu, Qing-Zhi Liu, Jie Long, Liang Li, Cai-Yue Gao, Shu-Han Yang, Zhi-Bin Zhao, and Zhe-Xiong Lian. Multi-omics analysis identifies FoxO1 as a regulator of macrophage function through

metabolic reprogramming. *Cell Death Disease*, 11(9):800, 2020. ISSN 2041-4889. doi: 10.1038/s41419-020-02982-0.

[244] Wuqiang Fan, Hidetaka Morinaga, Jane J. Kim, Eunju Bae, Nathanael J. Spann, Sven Heinz, Christopher K. Glass, and Jerrold M. Olefsky. FoxO1 regulates Tlr4 inflammatory pathway signalling in macrophages. *The EMBO Journal*, 29(24):4223–4236, 2010. doi: 10.1038/emboj.2010.268.

[245] David A. Hume. The Many Alternative Faces of Macrophage Activation. *Frontiers in Immunology*, 6(370), 2015. ISSN 1664-3224. doi: 10.3389/fimmu.2015.00370.

[246] Michael Rehli, Sabine Sulzbacher, Sabine Pape, Timothy Ravasi, Christine A Wells, Sven Heinz, Liane Söllner, Carol El Chartouni, Stefan W Krause, Eirikur Steingrimsson, David A Hume, and Reinhard Andreesen. Transcription factor Tfec contributes to the IL-4-inducible expression of a small group of genes in mouse macrophages including the granulocyte colony-stimulating factor receptor. *The Journal of Immunology*, 174 (11):7111–7122, 2005. ISSN 1550-6606. doi: 10.4049/jimmunol.174.11.7111.

[247] Xuan Su, Junzhi Tian, Binghua Li, Lixiao Zhou, Hui Kang, Zijie Pei, Mengyue Zhang, Chen Li, Mengqi Wu, Qian Wang, Bin Han, Chen Chu, Yaxian Pang, Jie Ning, Boyuan Zhang, Yujie Niu, and Rong Zhang. Ambient PM2.5 caused cardiac dysfunction through FoxO1-targeted cardiac hypertrophy and macrophage-activated fibrosis in mice. *Chemosphere*, 247(125881), 2020. ISSN 0045-6535. doi: 10.1016/j.chemosphere.2020.125881.

[248] Jill E. Moore, Michael J. Purcaro, Henry E. Pratt, Charles B. Epstein, Noam Shoresh, Jessika Adrian, Trupti Kawli, Carrie A. Davis, Alexander Dobin, Rajinder Kaul, Jessica Halow, Eric L. Van Nostrand, Peter Freese, David U. Gorkin, Yin Shen, Yupeng He, Mark Mackiewicz, Florencia Pauli-Behn, Brian A. Williams, Ali Mortazavi, Cheryl A. Keller, Xiao-Ou Zhang, Shaimae I. Elhajjajy, Jack Huey, Diane E. Dickel, Valentina Snetkova, Xintao Wei, Xiaofeng Wang, Juan Carlos Rivera-Mulia, Joel Rozowsky, Jing Zhang, Surya B. Chhetri, Jialing Zhang, Alec Victorsen, Kevin P. White, Axel Visel, Gene W. Yeo, Christopher B. Burge, Eric Lécuyer, David M. Gilbert, Job Dekker, John Rinn, Eric M. Mendenhall, Joseph R. Ecker, Manolis Kellis, Robert J. Klein, William S. Noble, Anshul Kundaje, Roderic Guigó, Peggy J. Farnham, J Michael Cherry, Richard M. Myers, Bing Ren, Brenton R. Graveley, Mark B. Gerstein, Len A. Pennacchio, Michael P. Snyder, Bradley E. Bernstein, Barbara Wold, Ross C. Hardison, Thomas R. Gingeras, John A. Stamatoyannopoulos, and Zhiping Weng. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, 583(7818): 699–710, 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2493-4.

[249] Jeffrey M. Granja, Ryan M. Corces, Sarah E. Pierce, S. Tansu Bagdatli, Hani Choudhry, Howard Y. Chang, and William J. Greenleaf. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nature Genetics*, 53 (3):403–411, 2021. ISSN 1546-1718. doi: 10.1038/s41588-021-00790-6.

[250] Sheng Wang, Yinlong Liao, Haoyuan Zhang, Yunqi Jiang, Zhelun Peng, Ruimin Ren, Xinyun Li, and Heng Wang. Tcf12 is required to sustain myogenic genes synergism with MyoD by remodelling the chromatin landscape. *Communications Biology*, 5(1): 1201, 2022. ISSN 2399-3642. doi: 10.1038/s42003-022-04176-0.

[251] Matthew J Potthoff, Hai Wu, Michael A Arnold, John M Shelton, Johannes Backs, John McAnally, James A Richardson, Rhonda Bassel-Duby, and Eric N. Olson. Histone

deacetylase degradation and MEF2 activation promote the formation of slow-twitch myofibers. *Journal of Clinical Investigation*, 117(9):2459–2467, 2007. ISSN 1558-8238. doi: 10.1172/JCI31960.

[252] Courtney M Anderson, Jianxin Hu, Ralston M Barnes, Analeah B Heidt, Ivo Cornelissen, and Brian L. Black. Myocyte enhancer factor 2C function in skeletal muscle is required for normal growth and glucose metabolism in mice. *Skeletal Muscle*, 27(5):7, 2015. ISSN 2044-5040. doi: 10.1186/s13395-015-0031-0.

[253] Alex Hennebry, Carole Berry, Victoria Siriett, Paul O'Callaghan, Linda Chau, Trevor Watson, Mridula Sharma, and Ravi Kambadur. Myostatin regulates fiber-type composition of skeletal muscle by regulating MEF2 and MyoD gene expression. *American Journal of Physiology-Cell Physiology*, 296(3):C525–C534, 2009. ISSN 1522-1563. doi: 10.1152/ajpcell.00259.2007.

[254] Hai Wu, Francisco J. Naya, Timothy A. McKinsey, Brian Mercer, John M. Shelton, Eva R. Chin, Alain R. Simard, Robin N. Michel, Rhonda Bassel-Duby, Eric N. Olson, and R. Sanders Williams. MEF2 responds to multiple calcium-regulated signals in the control of skeletal muscle fiber type. *EMBO Journal*, 19(9):1963–1973, 2000. ISSN 0261-4189. doi: 10.1093/emboj/19.9.1963.

[255] Jae-Hwan Jeong, Jung-Sook Jin, Hyun-Nam Kim, Sang-Min Kang, Julie C. Liu, Christopher J. Lengner, Florian Otto, Stefan Mundlos, Janet L. Stein, Andre J. van Wijnen, Jane B. Lian, Gary S. Stein, and Je-Yong Choi. Expression of Runx2 transcription factor in non-skeletal tissues, sperm and brain. *Journal of Cellular Physiology*, 217(2):511–517, 2008. doi: 10.1002/jcp.21524.

[256] Nadiya M. Teplyuk, Ying Zhang, Yang Lou, John R. Hawse, Mohammad Q. Hassan, Viktor I. Teplyuk, Jitesh Pratap, Mario Galindo, Janet L. Stein, Gary S. Stein, and Andre J. van Wijnen. The osteogenic transcription factor runx2 controls genes involved in sterol/steroid metabolism, including CYP11A1 in osteoblasts. *Molecular Endocrinology*, 23(6):849–861, 2009. ISSN 0888-8809. doi: 10.1210/me.2008-0270.

[257] Omar Khalid, Sanjeev K. Baniwal, Daniel J. Purcell, Nathalie Leclerc, Yankel Gabet, Michael R. Stallcup, Gerhard A. Coetzee, and Baruch Frenkel. Modulation of Runx2 Activity by Estrogen Receptor-: Implications for Osteoporosis and Breast Cancer. *Endocrinology*, 149(12):5984—-5995, 2008. doi: 10.1210/en.2008-0680.

[258] Anne-Louise Gannon, Laura O'Hara, J. Ian Mason, Anne Jørgensen, Hanne Frederiksen, Laura Milne, Sarah Smith, Rod T. Mitchell, and Lee B. Smith. Androgen receptor signalling in the male adrenal facilitates X-zone regression, cell turnover and protects against adrenal degeneration during ageing. *Scientific Reports*, 9(1):10457, 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-46049-3.

[259] Anne-Louise Gannon, Laura O'Hara, Ian J. Mason, Anne Jørgensen, Hanne Frederiksen, Michael Curley, Laura Milne, Sarah Smith, Rod T. Mitchell, and Lee B. Smith. Androgen Receptor Is Dispensable for X-Zone Regression in the Female Adrenal but Regulates Post-Partum Corticosterone Levels and Protects Cortex Integrity. *Frontiers in Endocrinology*, 11:599869, 2020. ISSN 1664-2392. doi: 10.3389/fendo.2020.599869.

[260] Kelly L. Kane, Chantal M. Longo-Guess, Leona H. Gagnon, Dalian Ding, Richard J. Salvi, and Kenneth R. Johnson. Genetic background effects on age-related hearing loss associated with Cdh23 variants in mice. *Hearing Research*, 283(1-2):80–88, 2012. ISSN 0378-5955. doi: 10.1016/j.heares.2011.11.007.

[261] Maximilian Zeidler, Alexander Hüttenhofer, Michaela Kress, and Kai K. Kummer. Intragenic MicroRNAs Autoregulate Their Host Genes in Both Direct and Indirect Ways-A Cross-Species Analysis. *Cells*, 9(1):232, 2020. ISSN 2073-4409. doi: 10.3390/cells9010232.

[262] Vladimir V. Galatenko, Alexey V. Galatenko, Timur R. Samatov, Andrey A. Turchinovich, Maxim Yu. Shkurnikov, Julia A. Makarova, and Alexander G. Tonevitsky. Comprehensive network of miRNA-induced intergenic interactions and a biological role of its core in cancer. *Scientific Reports*, 8(1):2418, 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-20215-5.

[263] Lyudmila F. Gulyaeva and Nicolay E. Kushlinskiy. Regulatory mechanisms of microRNA expression. *Journal of Translational Medicine*, 14(1):143, 2016. ISSN 1479-5876. doi: 10.1186/s12967-016-0893-x.

[264] Stephen J. Fleming, Mark D. Chaffin, Alessandro Arduini, Amer-Denis Akkad, Eric Banks, John C. Marioni, Anthony A. Philippakis, Patrick T. Ellinor, and Mehrtash Babadi. Unsupervised removal of systematic background noise from droplet-based single-cell experiments using CellBender. *Nature Methods*, 20(9):1323–1335, 2023. ISSN 1548-7091. doi: 10.1038/s41592-023-01943-7.

[265] Boris Muzellec, Maria Teleńczuk, Vincent Cabeli, and Mathieu Andreux. PyDESeq2: a python package for bulk RNA-seq differential expression analysis. *Bioinformatics*, 39(9):btad547, 2023. ISSN 1460-2059. doi: 10.1093/bioinformatics/btad547.

[266] Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J. Lennon, Kenneth J. Livak, Tarjei S. Mikkelsen, and John L. Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, 32(4):381–386, 2014. ISSN 1546-1696. doi: 10.1038/nbt.2859.

[267] Xiaojie Qiu, Qi Mao, Ying Tang, Li Wang, Raghav Chawla, Hannah A. Pliner, and Cole Trapnell. Reversed graph embedding resolves complex single-cell trajectories. *Nature Methods*, 14(10):979–982, 2017. ISSN 1548-7091. doi: 10.1038/nmeth.4402.

[268] Junyue Cao, Malte Spielmann, Xiaojie Qiu, Xingfan Huang, Daniel M. Ibrahim, Andrew J. Hill, Fan Zhang, Stefan Mundlos, Lena Christiansen, Frank J. Steemers, Cole Trapnell, and Jay Shendure. The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, 566(7745):496–502, 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-0969-x.

[269] Jacob H. Levine, Erin F. Simonds, Sean C. Bendall, Kara L. Davis, El-ad D. Amir, Michelle D. Tadmor, Oren Litvin, Harris G. Fienberg, Astraea Jager, Eli R. Zunder, Rachel Finck, Amanda L. Gedman, Ina Radtke, James R. Downing, Dana Pe'er, and Garry P. Nolan. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell*, 162(1):184–197, 2015. ISSN 1097-4172. doi: 10.1016/j.cell.2015.05.047.

[270] A. Sina Booeshaghi, Ingileif B. Hallgrímsdóttir, Ángel Gálvez-Merchán, and Lior Pachter. Depth normalization for single-cell genomics count data. *bioRxiv*, page 2022.05.06.490859, 2022. doi: 10.1101/2022.05.06.490859.

[271] Ieva Rauluseviciute, Rafael Riudavets-Puig, Romain Blanc-Mathieu, Jaime A Castro-Mondragon, Katalin Ferenc, Vipin Kumar, Roza Berhanu Lemma, Jérémy Lucas, Jeanne Chèneby, Damir Baranasic, Aziz Khan, Oriol Fornes, Sveinung Gundersen,

Morten Johansen, Eivind Hovig, Boris Lenhard, Albin Sandelin, Wyeth W Wasserman, François Parcy, and Anthony Mathelier. JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 52(D1):D174–D182, 2024. ISSN 1362-4962. doi: 10.1093/nar/gkad1059.

[272] Schep A. motifmatchr: Fast Motif Matching in R. 2023. doi: 10.18129/B9.bioc.motifmatchr.

[273] Michael Lawrence, Wolfgang Huber, Hervé Pagès, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin T. Morgan, and Vincent J. Carey. Software for computing and annotating genomic ranges. *PLOS Computational Biology*, 9(8):e1003118, 2013. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003118.

[274] Bin Li, Tao Qing, Jinhang Zhu, Zhuo Wen, Ying Yu, Ryutaro Fukumura, Yuanting Zheng, Yoichi Gondo, and Leming Shi. A Comprehensive Mouse Transcriptomic BodyMap across 17 Tissues by RNA-seq. *Scientific Reports*, 7(1):4200, 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-04520-z.

[275] Chien-Ping Ko and Richard Robitaille. Perisynaptic Schwann Cells at the Neuromuscular Synapse: Adaptable, Multitasking Glial Cells. *Cold Spring Harbor Perspectives in Biology*, 7(10):a020503, 2015. ISSN 1943-0264. doi: 10.1101/cshperspect.a020503.

[276] Aldrin K Y Yim, Peter L Wang, John R Bermingham Jr, Amber Hackett, Amy Strickland, Timothy M Miller, Cindy Ly, Robi D Mitra, and Jeffrey Milbrandt. Disentangling glial diversity in peripheral nerves at single-nuclei resolution. *Nature Neuroscience*, 25 (2):238–251, 2022. ISSN 1097-6256. doi: 10.1038/s41593-021-01005-1.

[277] Alison R Preston and Howard Eichenbaum. Interplay of hippocampus and prefrontal cortex in memory. *Current Biology*, 23(17):R764–R773, 2013. ISSN 0960-9822. doi: 10.1016/j.cub.2013.05.041.

[278] Inah Hwang, Heng Pan, Jun Yao, Olivier Elemento, Hongwu Zheng, and Jihye Paik. CIC is a critical regulator of neuronal differentiation. *JCI Insight*, 5(9):e135826, 2020. ISSN 2379-3708. doi: 10.1172/jci.insight.135826.

[279] Alice Grison, Zahra Karimaddini, Jeremie Breda, Tanzila Mukhtar, Marcelo Boareto, Katja Eschbach, Christian Beisel, Dagmar Iber, Erik van Nimwegen, Verdon Taylor, and Suzana Atanasoski. The protooncogene Ski regulates the neuron-glia switch during development of the mammalian cerebral cortex. *bioRxiv*, page 2022.12.16.520470, 2022. doi: 10.1101/2022.12.16.520470.

[280] Michael R Vogl, Simone Reiprich, Melanie Küspert, Thomas Kosian, Heinrich Schrewe, Klaus-Armin Nave, and Michael Wegner. Sox10 cooperates with the mediator subunit 12 during terminal differentiation of myelinating glia. *Journal of Neuroscience*, 33(15): 6679–6690, 2013. ISSN 1529-2401. doi: 10.1523/JNEUROSCI.5178-12.2013.

[281] Taasin Srivastava, Parham Diba, Justin M Dean, Fatima Banine, Daniel Shaver, Matthew Hagen, Xi Gong, Weiping Su, Ben Emery, Daniel L Marks, Edward N Harris, Bruce Baggenstoss, Paul H Weigel, Larry S Sherman, and Stephen A Back. A TLR/AKT/FoxO3 immune tolerance-like pathway disrupts the repair capacity of oligodendrocyte progenitors. *Journal of Clinical Investigation*, 128(5):2025–2041, 2018. ISSN 1558-8238. doi: 10.1172/JCI94158.

[282] Catherine Larochelle, Beatrice Wasser, Hélène Jamann, Julian T Löffel, Qiao-Ling Cui, Olivier Tastet, Miriam Schillner, Dirk Luchtman, Jérôme Birkenstock, Albrecht Stroh, Jack Antel, Stefan Bittner, and Frauke Zipp. Pro-inflammatory T helper 17 directly

harms oligodendrocytes in neuroinflammation. *Proceedings of the National Academy of Sciences of the United States of America*, 118(34):e2025813118, 2021. ISSN 1091-6490. doi: 10.1073/pnas.2025813118.

[283] S Martínez and L Puelles. Neurogenetic compartments of the mouse diencephalon and some characteristic gene expression patterns. *Results and Problems in Cell Differentiation*, 30:91–106, 2000. ISSN 1861-0412. doi: 10.1007/978-3-540-48002-0_4.

[284] K L Brunson, S Avishai-Eliner, C G Hatalski, and T Z Baram. Neurobiology of the stress response early in life: evolution of a concept and the role of corticotropin releasing hormone. *Molecular Psychiatry*, 6(6):647–656, 2001. ISSN 1476-5578. doi: 10.1038/sj.mp.4000942.

[285] Catherine J. Dunlavey. Introduction to the Hypothalamic-Pituitary-Adrenal Axis: Healthy and Dysregulated Stress Responses, Developmental Stress and Neurodegeneration. *Journal of Undergraduate Neuroscience*, 16(2):R59–R60, 2018. ISSN 1544-2896.

[286] A Acevedo-Rodriguez, A S Kauffman, B D Cherrington, C S Borges, T A Roepke, and M. Laconi. Emerging insights into hypothalamic-pituitary-gonadal axis regulation and interaction with stress signalling. *Journal of Neuroendocrinology*, 30(10):e12590, 2018. ISSN 1365-2826. doi: 10.1111/jne.12590.

[287] Guck T Ooi, Neveen Tawadros, and Ruth M. Escalona. Pituitary cell lines and their endocrine applications. *Molecular and Cellular Endocrinology*, 228(1-2):1–21, 2004. ISSN 0303-7207. doi: 10.1016/j.mce.2004.07.018.

[288] Aaron B. Lerner and Joseph S. McGuire. Melanocyte-Stimulating Hormone and Adrenocorticotrophic Hormone Their Relation to Pigmentation. *The New England Journal of Medicine*, 12(270):539–546, 1964. ISSN 1533-4406. doi: 10.1056/NEJM196403122701101.

[289] Erin M Wolf Horrell, Mary C Boulanger, and John A D'Orazio. Melanocortin 1 Receptor: Structure, Function, and Regulation. *Frontiers in Genetics*, 31(7):95, 2016. ISSN 1664-8021. doi: 10.3389/fgene.2016.00095.

[290] Yugong Ho, Peng Hu, Michael T Peel, Sixing Chen, Pablo G Camara, Douglas J Epstein, Hao Wu, and Stephen A. Liebhaber. Single-cell transcriptomic analysis of adult mouse pituitary reveals sexual dimorphism and physiologic demand-induced cellular plasticity. *Protein Cell*, 11(8):565–583, 2020. ISSN 1674-8018. doi: 10.1007/s13238-020-00705-x.

[291] Frederique Ruf-Zamojski, Zidong Zhang, Michel Zamojski, Gregory R Smith, Natalia Mendelev, Hanqing Liu, German Nudelman, Mika Moriwaki, Hanna Pincas, Rosa Gomez Castanon, Venugopalan D Nair, Nitish Seenarine, Mary Anne S Amper, Xiang Zhou, Luisina Ongaro, Chirine Toufaily, Gauthier Schang, Joseph R Nery, Anna Bartlett, Andrew Aldridge, Nimisha Jain, Gwen V Childs, Olga G Troyanskaya, Joseph R Ecker, Judith L Turgeon, Corrine K Welt, Daniel J Bernard, and Stuart C. Sealfon. Single nucleus multi-omics regulatory landscape of the murine pituitary. *Nature Communications*, 12(1):2677, 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-22859-w.

[292] Kai-Wen Ren, Xiao-Hong Yu, Yu-Hui Gu, Xin Xie, Yu Wang, Shi-Hao Wang, Hui-Hua Li, and Hai-Lian Bi. Cardiac-specific knockdown of Bhlhe40 attenuates angiotensin II (Ang II)-Induced atrial fibrillation in mice. *Frontiers in Cardiovascular Medicine*, 11 (9):957903, 2022. ISSN 2297-055X. doi: 10.3389/fcvm.2022.957903.

[293] M. E. Young, P. Razeghi, and H. Taegtmeyer. Clock genes in the heart: characterization and attenuation with hypertrophy. *Circulation Research*, 88(11):1142–1150, 2001. ISSN 0009-7330. doi: 10.1161/hh1101.091190.

[294] Peng Jiang, Kathleen M Franklin, Marilyn J Duncan, Bruce F O'Hara, and Jonathan P. Wisor. Distinct phase relationships between suprachiasmatic molecular rhythms, cerebral cortex molecular rhythms, and behavioral rhythms in early runner (CAST/EiJ) and nocturnal (C57BL/6J) mice. *Sleep*, 35(10):1385–1394, 2012. ISSN 1550-9109. doi: 10.5665/sleep.2120.

[295] Gavin P. Vinson. Functional Zonation of the Adult Mammalian Adrenal Cortex. *Frontiers in Neuroscience*, 2016. ISSN 1662-453X. doi: 10.3389/fnins.2016.00238.

[296] Dona Lee Wong. Epinephrine biosynthesis: hormonal and neural control during stress. *Cellular and Molecular Neurobiology*, 26(4-6):891–900, 2006. ISSN 1573-6830. doi: 10.1007/s10571-006-9056-6.

[297] Junhua Zhou, Lalarukh Haris Shaikh, Sudeshna G Neogi, Ian McFarlane, Wanfeng Zhao, Nichola Figg, Cheryl A Brighton, Carmela Maniero, Ada E D Teo, Elena A B Azizan, and Morris J. Brown. DACH1, a zona glomerulosa selective gene in the human adrenal, activates transforming growth factor- signaling and suppresses aldosterone secretion. *Hypertension*, 65(5):1103–1110, 2015. ISSN 1473-5598. doi: 10.1161/HYP.0000000000000025.

[298] Kelly L Sams, Chinatsu Mukai, Brooke A Marks, Chitvan Mittal, Elena Alina Demeter, Sophie Nelissen, Jennifer K Grenier, Ann E Tate, Faraz Ahmed, and Scott A. Coonrod. Delayed puberty, gonadotropin abnormalities and subfertility in male Padi2/Padi4 double knockout mice. *Reproductive Biology and Endocrinology*, 20(1):150, 2022. ISSN 1477-7827. doi: 10.1186/s12958-022-01018-w.

[299] Jia L Zhuo and Xiao C. Li. Proximal nephron. *Comprehensive Physiology*, 3(3):1079–1123, 2013. ISSN 2040-4603. doi: 10.1002/cphy.c110061.

[300] Andrew Ransick, Nils O Lindström, Jing Liu, Qin Zhu, Jin-Jin Guo, Gregory F. Alvarado, Albert D. Kim, Hannah G. Black, Junhyong Kim, and Andrew P. McMahon. Single-Cell Profiling Reveals Sex, Lineage, and Regional Diversity in the Mouse Kidney. *Developmental Cell*, 51(3):399–413, 2019. ISSN 1534-5807. doi: 10.1016/j.devcel.2019.10.005.

[301] H Pavenstädt. Roles of the podocyte in glomerular function. *American Journal of Physiology-Renal Physiology*, 278(2):F173–F179, 2000. ISSN 1931-857X. doi: 10.1152/ajprenal.2000.278.2.F173.

[302] Thomas L. Pannabecker. Structure and function of the thin limbs of the loop of Henle. *Comprehensive Physiology*, 2(3):2063–2086, 2012. ISSN 2040-4603. doi: 10.1002/cphy.c110019.

[303] Zehao Zhang, Chloe Schaefer, Weirong Jiang, Ziyu Lu, Jasper Lee, Andras Sziraki, Abdulraouf Abdulraouf, Brittney Wick, Maximilian Haeussler, Zhuoyan Li, Gesmira Molla, Rahul Satija, Wei Zhou, and Junyue Cao. A Panoramic View of Cell Population Dynamics in Mammalian Aging. *bioRxiv*, page 2024.03.01.583001, 2024. doi: 10.1101/2024.03.01.583001.

[304] Zhi-Hang Li, Xiao-Yan Guo, Xiao-Ying Quan, Chen Yang, Ze-Jian Liu, Hong-Yong Su, Ning An, and Hua-Feng Liu. The Role of Parietal Epithelial Cells in the Pathogenesis of Podocytopathy. *Frontiers in Physiology*, 11(13):832772, 2022. ISSN 1664-042X. doi:

10.3389/fphys.2022.832772.

[305] Ville Veikkolainen, Florence Naillat, Antti Railo, Lijun Chi, Aki Manninen, Peter Hohenstein, Nick Hastie, Seppo Vainio, and Klaus Elenius. ErbB4 modulates tubular cell polarity and lumen diameter during kidney development. *Journal of the American Society of Nephrology*, 23(1):112–122, 2012. ISSN 1533-3450. doi: 10.1681/ASN.2011020160.

[306] Qi Cai, Natalia I. Dmitrieva, Joan D. Ferraris, Heddwen L. Brooks, Bas W M. van Balkom, and Maurice Burg. Pax2 expression occurs in renal medullary epithelial cells in vivo and in cell culture, is osmoregulated, and promotes osmotic tolerance. *Proceedings of the National Academy of Sciences of the United States of America*, 102(2):503–508, 2005. ISSN 1091-6490. doi: 10.1073/pnas.0408840102.

[307] Elżbieta Złowocka-Perłowska, Aleksandra Tołoczko-Grabarek, Steven A Narod, and Jan Lubiński. Germline BRCA1 and BRCA2 mutations and the risk of bladder or kidney cancer in Poland. *Hereditary Cancer in Clinical Practice*, 20(1):13, 2022. ISSN 1897-4287. doi: 10.1186/s13053-022-00220-6.

[308] Huayi Feng, Shouqing Cao, Qing Ouyang, Huaikang Li, Xiubin Li, Ke Chen, Xiangyi Zhang, Yan Huang, Xu Zhang, and Xin Ma. Prevalence of germline mutations in cancer susceptibility genes in Chinese patients with renal cell carcinoma. *Translational Andrology and Urology*, 12(2):308–319, 2023. ISSN 2223-4691. doi: 10.21037/tau-23-32.

[309] Maria João Ferreira, Ana Sílvia Pires-Luís, Márcia Vieira-Coimbra, Pedro Costa-Pinheiro, Luís Antunes, Paula C. Dias, Francisco Lobo, Jorge Oliveira, Céline S. Gonçalves, Bruno M. Costa, Rui Henrique, and Carmen Jerónimo. SETDB2 and RIOX2 are differentially expressed among renal cell tumor subtypes, associating with prognosis and metastization. *Epigenetics*, 12(12):1057–1064, 2017. ISSN 1559-2294. doi: 10.1080/15592294.2017.1385685.

[310] Shani Ben-Moshe and Shalev Itzkovitz. Spatial heterogeneity in the mammalian liver. *Nature Reviews Gastroenterology Hepatology*, 16(7):395–410, 2019. ISSN 1759-5045. doi: 10.1038/s41575-019-0134-x.

[311] Janie L Baratta, Anthony Ngo, Bryan Lopez, Natasha Kasabwalla, Kenneth J Longmuir, and Richard T Robertson. Cellular organization of normal mouse liver: a histological, quantitative immunocytochemical, and fine structural analysis. *Histochemistry and Cell Biology*, 131(6):713–726, 2009. ISSN 0948-6143. doi: 10.1007/s00418-009-0577-1.

[312] Avner Ehrlich, Daniel Duche, Gladys Ouedraogo, and Yaakov Nahmias. Challenges and Opportunities in the Design of Liver-on-Chip Microdevices. *Annual Review of Biomedical Engineering*, 4(21):219–239, 2019. ISSN 1545-4274. doi: 10.1146/annurev-bioeng-060418-052305.

[313] James L Boyer. Bile formation and secretion. *Comprehensive Physiology*, 3(3):1035–1078, 2013. ISSN 2040-4603. doi: 10.1002/cphy.c120027.

[314] Laura J Dixon, Mark Barnes, Hui Tang, Michele T Pritchard, and Laura E Nagy. Kupffer cells in the liver. *Comprehensive Physiology*, 3(2):785–797, 2013. ISSN 2040-4603. doi: 10.1002/cphy.c120026.

[315] Marc Y Donath and Steven E Shoelson. Type 2 diabetes as an inflammatory disease. *Nature Reviews Immunology*, 11(2):98–107., 2011. ISSN 1474-1741. doi: 10.1038/nri2925.

[316] Tara L Conforto, Yijing Zhang, Jennifer Sherman, and David J Waxman. Impact of CUX2 on the female mouse liver transcriptome: activation of female-biased genes and repression of male-biased genes. *Molecular Biology of the Cell*, 32(22):4611–4627, 2012. ISSN 1939-4586. doi: 10.1128/MCB.00886-12.

[317] Joni Nikkanen, Yew Ann Leong, William C Krause, Denis Dermadi, J Alan Maschek, Tyler Van Ry, James E Cox, Ethan J Weiss, Omer Gokcumen, Ajay Chawla, and Holly A Ingraham. An evolutionary trade-off between host immunity and metabolism drives fatty liver in male mice. *Science*, 378(6617):290–295, 2022. ISSN 1095-9203. doi: 10.1126/science.abn9886.

[318] Weiling Zheng, Hongyan Xu, Siew Hong Lam, Huaien Luo, R Krishna Murthy Karuturi, and Zhiyuan Gong. Transcriptomic analyses of sexual dimorphism of the zebrafish liver and the effect of sex hormones. *PLOS One*, 8(1):e53562, 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0053562.

[319] Xiaomin Hong, Sanqiang Li, Renli Luo, Mengli Yang, Junfei Wu, Shuning Chen, and Siyu Zhu. Mechanisms of the TGF-1/Smad3-signaling pathway in gender differences in alcoholic liver fibrosis. *The Journal of Physiological Sciences*, 74(1):13, 2024. ISSN 1880-6562. doi: 10.1186/s12576-024-00901-y.

[320] Sang Gyun Noh, Hee Jin Jung, Seungwoo Kim, Radha Arulkumar, Dae Hyun Kim, Daeui Park, and Hae Young Chung. Regulation of Circadian Genes Nr1d1 and Nr1d2 in Sex-Different Manners during Liver Aging. *International Journal of Molecular Sciences*, 23(17):10032, 2022. ISSN 1422-0067. doi: 10.3390/ijms231710032.

[321] Monika Gjorgjieva, Anne-Sophie Ay, Marta Correia de Sousa, Etienne Delangre, Dobrochna Dolicka, Cyril Sobolewski, Christine Maeder, Margot Fournier, Christine Sempoux, and Michelangelo Foti. MiR-22 Deficiency Fosters Hepatocellular Carcinoma Development in Fatty Liver. *Cells*, 11(18):2860, 2022. ISSN 2073-4409. doi: 10.3390/cells11182860.

[322] Eva Moreno, Mathilda J M Toussaint, Saskia C van Essen, Laura Bongiovanni, Elsbeth A van Liere, Mirjam H Koster, Ruixue Yuan, Jan M van Deursen, Bart Westendorp, and Alain de Bruin. E2F7 Is a Potent Inhibitor of Liver Tumor Growth in Adult Mice. *Hepatology*, 73(1):303–317, 2021. ISSN 1527-3350. doi: 10.1002/hep.31259.

[323] Keisuke Yoshida, Toshio Maekawa, Nhung Hong Ly, Shin-Ichiro Fujita, Masafumi Muratani, Minami Ando, Yuki Katou, Hiromitsu Araki, Fumihito Miura, Katsuhiko Shirahige, Mariko Okada, Takashi Ito, Bruno Chatton, and Shunsuke Ishii. ATF7-Dependent Epigenetic Changes Are Required for the Intergenerational Effect of a Paternal Low-Protein Diet. *Molecular Cell*, 78(3):445–458, 2020. ISSN 1097-2765. doi: 10.1016/j.molcel.2020.02.028.

[324] R. Sharma and A Agarwal. *Spermatogenesis: An Overview*. Springer, New York, NY, 2011. ISBN 978-1-4419-6857-9. doi: 10.1007/978-1-4419-6857-9_2.

[325] Theodore R Chauvin and Michael D. Griswold. Androgen-regulated genes in the murine epididymis. *Biology of Reproduction*, 71(2):560–569, 2004. ISSN 1529-7268. doi: 10.1095/biolreprod.103.026302.

[326] D M Kelly and T H Jones. Testosterone and obesity. *Obesity Reviews*, 16(7):581–606, 2015. ISSN 1467-789X. doi: 10.1111/obr.12282.

[327] Roland Kotolloshi, Mieczyslaw Gajda, Marc-Oliver Grimm, and Daniel Steinbach. Wnt/-Catenin Signalling and Its Cofactor BCL9L Have an Oncogenic Effect in Bladder

Cancer Cells. *International Journal of Molecular Sciences*, 23(10):5319, 2022. ISSN 1422-0067. doi: 10.3390/ijms23105319.

[328] Xiaohui Jiang, Xiang Wang, Xueguang Zhang, Zhun Xiao, Chaoliang Zhang, Xiaojun Liu, Jinyan Xu, Dingming Li, and Ying Shen. A homozygous RNF220 mutation leads to male infertility with small-headed sperm. *Gene*, 10(688):13–18, 2019. ISSN 2073-4425. doi: 10.1016/j.gene.2018.11.074.

[329] Cailin Wilson and Adam J. Krieg. KDM4B: A Nail for Every Hammer? *Genes*, 10 (2):134, 2019. ISSN 2073-4425. doi: 10.3390/genes10020134.

[330] Mary E Morris, Marie-Charlotte Meinsohn, Maeva Chauvin, Hatice D Saatcioglu, Aki Kashiwagi, Natalie A Sicher, Ngoc Nguyen, Selena Yuan, Rhian Stavely, Minsuk Hyun, Patricia K Donahoe, Bernardo L Sabatini, and David Pépin. A single-cell atlas of the cycling murine ovary. *eLife*, 7(11):e77239, 2022. ISSN 2050-084X. doi: 10.7554/eLife.77239.

[331] Ting Liu, Yifei Huang, and Hui Lin. Estrogen disorders: Interpreting the abnormal regulation of aromatase in granulosa cells. *International Journal of Molecular Medicine*, 47(5):73, 2021. ISSN 1791-244X. doi: 10.3892/ijmm.2021.4906.

[332] C Allison Stewart and Richard R Behringer. Mouse oviduct development. *Results and Problems in Cell Differentiation*, 55:247–262, 2012. ISSN 1861-0412. doi: 10.1007/978-3-642-30406-4_14.

[333] Y Huang, S-H Liang, L-B Xiang, X-T Han, W Zhang, J Tang, X-H Wu, and M-Q Zhang. miR-218 Promoted the Apoptosis of Human Ovarian Carcinoma Cells via Suppression of the WNT/-Catenin Signaling Pathway. *Molecular Biology*, 51(4):629–636, 2017. ISSN 1608-3245. doi: 10.7868/S0026898417030065.

[334] Rachel E Dickinson, Michelle Myers, and W Colin Duncan. Novel regulated expression of the SLIT/ROBO pathway in the ovary: possible role during luteolysis in women. *Endocrinology*, 149(10):5024–5034, 2008. ISSN 1479-6805. doi: 10.1210/en.2008-0204.

[335] Randy L. Bogan, Melinda J. Murphy, Richard L. Stouffer, and Jon D. Hennebold. Prostaglandin Synthesis, Metabolism, and Signaling Potential in the Rhesus Macaque Corpus Luteum throughout the Luteal Phase of the Menstrual Cycle. *Endocrinology*, 149(11):5861–5871, 2008. ISSN 1479-6805. doi: 10.1210/en.2008-0500.

[336] L. Devoto, P. Kohen, R. R. Gonzalez, O. Castro, I. Retamales, M. Vega, P. Carvallo, L. K. Christenson, and J. F. 3rd Strauss. Expression of steroidogenic acute regulatory protein in the human corpus luteum throughout the luteal phase. *The Journal of Clinical Endocrinology Metabolism*, 86(11):5633–5639, 2001. ISSN 1945-7197. doi: 10.1210/jcem.86.11.7982.

[337] Anne Bachelot and Nadine Binart. Corpus luteum development: lessons from genetic models in mice. *Current Topics in Developmental Biology*, 68:49–84, 2005. ISSN 0070-2153. doi: 10.1016/S0070-2153(05)68003-9.

[338] Jared Talbot and Lisa Maves. Skeletal muscle fiber type: using insights from muscle developmental biology to dissect targets for susceptibility and resistance to muscle disease. *Wiley Interdisciplinary Reviews: Developmental Biology*, 5(4):518–534, 2016. ISSN 1759-7692. doi: 10.1002/wdev.230.

[339] Camilla Reina Maroni, Michael A Friedman, Yue Zhang, Michael J McClure, Stefania Fulle, Charles R Farber, and Henry J. Donahue. Genetic variability affects the response of skeletal muscle to disuse. *Journal of Musculoskeletal and Neuronal Interactions*, 21

(3):387–396, 2021. ISSN 1108-7161.

[340] Margherita Ciano, Giada Mantellato, Martin Connolly, Mark Paul-Clark, Saffron Willis-Owen, Miriam F Moffatt, William O C M Cookson, Jane A Mitchell, Michael I Polkey, Simon M Hughes, Paul R Kemp, and S Amanda Natanek. EGF receptor (EGFR) inhibition promotes a slow-twitch oxidative, over a fast-twitch, muscle phenotype. *Scientific Reports*, 9(1):9218, 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-45567-4.

[341] Kohei Hamanaka, Darina Šikrová, Satomi Mitsuhashi, Hiroki Masuda, Yukari Sekiguchi, Atsuhiko Sugiyama, Kazumoto Shibuya, Richard J L F Lemmers, Remko Goossens, Megumu Ogawa, Koji Nagao, Chikashi Obuse, Satoru Noguchi, Yukiko K Hayashi, Satoshi Kuwabara, Judit Balog, Ichizo Nishino, and Silvère M. van der Maarel. Homozygous nonsense variant in LRIF1 associated with facioscapulohumeral muscular dystrophy. *Neurology*, 94(23):e2441–e2447, 2020. ISSN 1432-1459. doi: 10.1212/WNL.0000000000009617.

[342] Bo Hoon Lee, Phillip Mongiovi, Thierry Levade, Bethany Marston, Joan Mountain, and Emma Ciafaloni. Spinal muscular atrophy and Farber disease due to ASAH1 variants: A case report. *American Journal of Medical Genetics*, 182(10):2369–2371, 2020. ISSN 1552-4833. doi: 10.1002/ajmg.a.61764.

[343] Noemí De Luna, Eduard Gallardo, and Isabel Illa. In vivo and in vitro dysferlin expression in human muscle satellite cells. *Journal of Neuropathology Experimental Neurology*, 63(10):1104–1113, 2004. ISSN 1554-6578. doi: 10.1093/jnen/63.10.1104.

[344] Martin Horak, Jan Novak, and Julie Bienertova-Vasku. Muscle-specific microRNAs in skeletal muscle development. *Developmental Biology*, 410(1):1–13, 2016. ISSN 1095-564X. doi: 10.1016/j.ydbio.2015.12.013.

[345] Michael V Taylor and Simon M Hughes. Mef2 and the skeletal muscle differentiation program. *Seminars in Cell Developmental Biology*, 72:33–44, 2017. ISSN 1096-3634. doi: 10.1016/j.semcdb.2017.11.020.

[346] Luisa Boldrin, Peter S Zammit, and Jennifer E Morgan. Satellite cells from dystrophic muscle retain regenerative capacity. *Stem Cell Research*, 14(1):20–29, 2015. ISSN 1873-5061. doi: 10.1016/j.scr.2014.10.007.

[347] Adam Gayoso, Romain Lopez, Galen Xing, Pierre Boyeau, Valeh Valiollah Pour Amiri, Justin Hong, Katherine Wu, Michael Jayasuriya, Edouard Mehlman, Maxime Langevin, Yining Liu, Jules Samaran, Gabriel Misrachi, Achille Nazaret, Oscar Clivio, Chenling Xu, Tal Ashuach, Mariano Gabitto, Mohammad Lotfollahi, Valentine Svensson, Eduardo da Veiga Beltrame, Vitalii Kleshchevnikov, Carlos Talavera-López, Lior Pachter, Fabian J. Theis, Aaron Streets, Michael I. Jordan, Jeffrey Regier, and Nir Yosef. A Python library for probabilistic analysis of single-cell omics data. *Nature Biotechnology*, 40(2):163–166, 2022. ISSN 1087-0156. doi: 10.1038/s41587-021-01206-w.

[348] Chenling Xu, Romain Lopez, Edouard Mehlman, Jeffrey Regier, Michael I. Jordan, and Nir Yosef. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Molecular Systems Biology*, 17(1):e9620, 2021. ISSN 1744-4292. doi: 10.15252/msb.20209620.

[349] Alexander D. Diehl, Terrence F. Meehan, Yvonne M. Bradford, Matthew H. Brush, Wasila M. Dahdul, David S. Dougall, Yongqun He, David Osumi-Sutherland, Alan

Ruttenberg, Sirarat Sarntivijai, Ceri E. Van Slyke, Nicole A. Vasilevsky, Melissa A. Haendel, Judith A. Blake, and Christopher J. Mungall. The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. *Journal of Biomedical Semantics*, 7(1):44, 2016. ISSN 2041-1480. doi: 10.1186/s13326-016-0088-7.

[350] Yingyao Zhou, Bin Zhou, Lars Pache, Max Chang, Alireza Hadj Khodabakhshi, Olga Tanaseichuk, Christopher Benner, and Sumit K. Chanda. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nature Communications*, 10(1):1523, 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-09234-6.

[351] Anneleen Remmerie, Liesbet Martens, Tinne Thoné, Angela Castoldi, Ruth Seurinck, Benjamin Pavie, Joris Roels, Bavo Vanneste, Sofie De Prijck, Mathias Vanhockerhout, Mushida Binte Abdul Latib, Lindsey Devisscher, Anne Hoorens, Johnny Bonnardel, Niels Vandamme, Anna Kremer, Peter Borghgraef, Hans Van Vlierberghe, Saskia Lippens, Edward Pearce, Yvan Saeys, and Charlotte L. Scott. Osteopontin Expression Identifies a Subset of Recruited Macrophages Distinct from Kupffer Cells in the Fatty Liver. *Immunity*, 53(3):641–657, 2020. ISSN 1097-4180. doi: 10.1016/j.immuni.2020.08.004.

[352] Claudio Novella-Rausell, Magda Grudniewska, Dorien J M Peters, and Ahmed Mahfouz. A comprehensive mouse kidney atlas enables rare cell population characterization and robust marker discovery. *iScience*, 26(6):106877, 2023. ISSN 2589-0042. doi: 10.1016/j.isci.2023.106877.

[353] Karan Bedi, Brian R. Magnuson, Ishwarya Narayanan, Michelle Paulsen, Thomas E. Wilson, and Mats Ljungman. Co-transcriptional splicing efficiencies differ within genes and between cell types. *RNA*, 27(7):829–840, 2021. ISSN 1469-9001. doi: 10.1261/rna.078662.120.

[354] Elena Nikonova, Shao-Yen Kao, and Maria L. Spletter. Contributions of alternative splicing to muscle type development and function. *Seminars in Cell Developmental Biology*, 104:65–80, 2020. ISSN 1096-3634. doi: 10.1016/j.semcdb.2020.02.003.

[355] Simon A Hardwick, Wen Hu, Anoushka Joglekar, Li Fan, Paul G Collier, Careen Foord, Jennifer Balacco, Samantha Lanjewar, Maureen McGuirk Sampson, Frank Koopmans, Andrey D Prjibelski, Alla Mikheenko, Natan Belchikov, Julien Jarroux, Anne Bergstrom Lucas, Miklós Palkovits, Wenjie Luo, Teresa A Milner, Lishomwa C Ndhlovu, August B Smit, John Q Trojanowski, Virginia M Y Lee, Olivier Fedrigo, Steven A Sloan, Dóra Tombácz, M Elizabeth Ross, Erich Jarvis, Zsolt Boldogkői, Li Gan, and Hagen U. Tilgner. Single-nuclei isoform RNA sequencing unlocks barcoded exon connectivity in frozen brain tissue. *Nature Biotechnology*, 40(7):1082–1092, 2022. ISSN 1087-0156. doi: 10.1038/s41587-022-01231-3.

[356] Ming-Yi Chou, Dhivya Appan, Kai-Wei Chang, Chih-Hsuan Chou, Chia-Yi Lin, Susan Shur-Fen Gau, and Hsien-Sung Huang. Mouse hybrid genome mediates diverse brain phenotypes with the specificity of reciprocal crosses. *Federation of American Societies for Experimental Biology*, 36(3):e22232, 2022. ISSN 1530-6860. doi: 10.1096/fj.202101624R.

[357] Patricia J. Wittkopp, Belinda K. Haerum, and Andrew G. Clark. Evolutionary changes in cis and trans gene regulation. *Nature*, 430(6995):85–88, 2004. ISSN 0028-0836. doi: 10.1038/nature02698.

[358] Angela Goncalves, Sarah Leigh-Brown, David Thybert, Klara Stefflova, Ernest Turro,

Paul Flicek, Alvis Brazma, Duncan T Odom, and John C. Marioni. Extensive compensatory cis-trans regulation in the evolution of mouse gene expression. *Genome Research*, 22(12):2376–2384, 2012. ISSN 1088-9051. doi: 10.1101/gr.142281.112.

[359] K. R. Johnson, L. C. Erway, S. A. Cook, J. F. Willott, and Q. Y. Zheng. A major gene affecting age-related hearing loss in C57BL/6J mice. *Hearing Research*, 114(1-2): 83–92, 1997. ISSN 0378-5955. doi: 10.1016/s0378-5955(97)00155-x.

[360] R. Mott, C. J. Talbot, M. G. Turri, A. C. Collins, and J. Flint. A method for fine mapping quantitative trait loci in outbred animal stocks. *Proceedings of the National Academy of Sciences of the United States of America*, 97(23):12649–12654, 2000. ISSN 1091-6490. doi: 10.1073/pnas.230304397.

[361] M. G. P. van der Wijst, D. H. de Vries, H. E. Groot, G. Trynka, C. C. Hon, M. J. Bonder, O. Stegle, M. C. Nawijn, Y. Idaghdour, P. van der Harst, C. J. Ye, J. Powell, F. J. Theis, A. Mahfouz, M. Heinig, and L. Franke. The single-cell eQTLGen consortium. *eLife*, 9(9):e52155, 2020. ISSN 2050-084X. doi: 10.7554/eLife.52155.