Epigenomic and Transcriptomic Profiling for the Study of Monogenic and Polygenic Traits and Disease

by

Peter Orchard

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2020

Doctoral Committee:

Associate Professor Stephen C. J. Parker, Chair
Professor Charles F. Burant
Associate Professor Hyun Min Kang
Professor Jun Z. Li
Assistant Professor Jonathan Terhorst

Peter Orchard

porchard@umich.edu

ORCID iD: 0000-0001-6097-1106

To my wife, parents, and brothers

ACKNOWLEDGEMENTS

I have been fortunate to spend the last five years surrounded by knowledgeable, caring, and fun individuals. My thesis advisor Stephen Parker takes his role as a mentor seriously and always has his trainees' needs in mind. He has also made the lab environment a priority, and as a result I have benefited not only from Steve's quality mentorship but also from my interactions with my meticulous and kind lab mates. Dr. Ricardo Albanus, Dr. Arushi Varshney, and Vivek Rai were generous in sharing their experience and knowledge, scientific and otherwise. I also benefited greatly from observing John Hensley's work during my first few years in the lab, and I thank him for his patience and willingness to help (as well as working with me on the ataqv project discussed in this thesis). I collaborated on numerous experiments with Dr. Nandini Manickam and Dr. Yoshi Kyono who both took great pains to produce excellent work – their collective handiwork spans most of the chapters in this thesis. Anya Kiseleva, who I only briefly overlapped with in the lab, also produced some of the data presented here. Though I did not do much work with them directly, Dr. Venkat Elangovan, Dr. Danial Quang, Cynthia Zajac, and Christa Ventresca were supportive lab mates as well.

The Genome Sciences Training Program (GSTP) at the University provided me with both training opportunities and funding for three years. I want to thank the program directory Prof. Mike Boehnke and the program administrator Dawn Keane for their dedication to the GSTP and their hard work to make the training program fun and valuable. I also with to thank the National Human Genome Research Institute of the National Institutes of Health, which funds the GSTP (grant T32 HG00040).

I was also funded for one year by a University of Michigan Rackham Predoctoral Fellowship, for which I wish to thank the Rackham Graduate School.

Prof. Margit Burmeister and Prof. Maureen Sartor, the co-directors of the Bioinformatics Ph.D. program, have both been a great help to me and have been deliberate in their efforts to continually improve the Ph.D. program, to my own benefit as well as that of all the other students. I thank them for their guidance and advice. I would also like to thank the DCMB staff, most notably Julia Eussen, the Ph.D. program coordinator, and Jane Wiesner, who contribute substantially to daily student life in the Ph.D. program.

My thesis committee members – Professors Charles Burant, Jun Li, Jonathan Terhorst, and Hyun Min Kang, in addition to Steve – have always been willing to help and provide advice, and have been supportive in all my interactions with them.

The positive experiences and outstanding mentorship I received in my Master's thesis lab at my previous institution – Prof. Wolfgang Enard's lab at LMU Munich – played an outsize role in convincing me to pursue a Ph.D. and preparing me to succeed in a Ph.D. program. In addition to Wolfgang I wish to thank my direct supervisor Dr. Ines Hellmann especially, for helping me develop my technical abilities and her academic and career advice.

My parents, Michael and Susan, certainly deserve my gratitude as well. They have truly been supportive in every conceivable way as I've tried to figure out what exactly I want to do with my life. Lastly and most importantly I wish to thank my wife Małgosia. She has been incredibly patient and has sought to make my life better every day, despite all I've asked her to sacrifice so that I could pursue my Ph.D. at the University of Michigan. I hope I've been half the spouse to her that she's been to me, and I look forward to continuing the next phase of our journey together.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

Many trait-associated genomic loci are in non-coding regions of the genome. Determining which genetic variants in these regions are causally related to a trait and elucidating their downstream effects can be difficult. Layering transcriptomic and epigenomic data on top of genetic variation data can help nominate causal phenotype-associated variants and generate hypotheses about their effects in different cellular contexts.

In this thesis, I first apply RNA-sequencing (RNA-seq) and the assay for transposase accessible chromatin using sequencing (ATAC-seq) to investigate gene expression and chromatin accessibility in the Danforth mouse, a model of caudal birth defects. The Danforth phenotype results from an endogenous retroviral insertion near the *Ptf1a* gene. I identify 49 genes differentially expressed between Danforth and WT E9.5 tailbuds, including increased expression of *Ptf1a* and the nearby *Gm13344* lncRNA in Danforth. A gene ontology enrichment analysis indicates differentially expressed genes are enriched in the hedgehog signaling pathway, suggesting disruption of hedgehog signaling may cause the Danforth phenotype. I identify one region of increased chromatin accessibility in Danforth relative to WT mice, localizing to the *Gm13344* promoter. This region is orthologous to a human *PTF1A* enhancer, suggesting it may mediate *Ptf1a* overexpression in the Danforth mouse.

Next, I apply a software package for the quality control of ATAC-seq data (developed in our lab) to public datasets to measure heterogeneity, and analyze GM12878

ATAC-seq data to quantify the impact of Tn5 transposase concentration and sequencing lane cluster density. I find that increasing cluster density shifts the ATAC-seq fragment length distribution towards shorter fragments and results in greater transcription start site enrichment. I show that increasing Tn5 transposase concentration increases the enrichment of reads in enhancers and promoters, with 80% of ATAC-seq peaks showing increased signal with increasing Tn5 concentration (5% FDR). Peaks bound by the CTCF transcription factor are less sensitive to Tn5 concentration than those bound by other transcription factors. This analysis demonstrates the difficulties in reliably quantifying chromatin accessibility and utilizing public datasets.

I then apply single-nucleus ATAC-seq and RNA-seq to human and rat skeletal muscle to generate cell type specific transcriptomic and chromatin accessibility maps. I integrate these maps with UK Biobank genome-wide association study (GWAS) data to explore enrichment of GWAS signals in cell type specific ATAC-seq peaks. I demonstrate the utility of these maps by nominating causal genetic variants and cell types at several GWAS loci, including the T2D-associated *ARL15* locus. At the *ARL15* locus I nominate a credible set variant in a highly mesenchymal stem cell specific ATAC-seq peak.

Lastly, to gain insight into the genetic regulation of chromatin architecture and its association with aerobic exercise capacity, I analyze skeletal muscle ATAC-seq (n = 129) and RNA-seq (n = 143) from a rat model for untrained running capacity. Although no genes associate with running capacity at 5% FDR, a gene ontology enrichment analysis indicates that the genes with the strongest association are enriched in fatty acid oxidation pathways, consistent with previous findings in this rat model. I identify no ATAC-seq peaks associated with running capacity (5% FDR) but find 4,477 ATAC-seq peaks associate with at least one SNP (5% FDR).

Together, these projects demonstrate the value of epigenomic and transcriptomic data in the investigation of monogenic and polygenic traits, as well as the challenges

and limitations of applying epigenomic and transcriptomic data in this context.

CHAPTER I

Introduction

Many human diseases have a clear genetic component [218], and understanding the genetic basis of disease is one of the primary goals of biomedical research. By deciphering which genetic variants confer susceptibility, and why they associate with that disease, investigators hope to gain a better understanding of the pathophysiology of the disease, be able to predict which individuals will be most susceptible to the disease, and determine how the disease might be treated.

1.1 Non-coding genetic variation in common polygenic traits and disease

Recent progress in genotyping [93] and sequencing technologies [89] has substantially eased the task of associating genetic variation with phenotypic variation and disease. Dense genotyping chips enable accurate genotyping of large numbers of individuals, thereby enabling genetic variant allele frequencies to be associated with binary and continuous traits.

The favored modern method for studying the association between genetic variation and common, polygenic traits is the genome-wide association study (GWAS). A GWAS tests for associations between genetic variants across the genome and a given phenotype. Modern GWAS may involve hundreds of thousands of individuals and tens of millions of genetic variants (generally single-nucleotide polymorphisms (SNPs), which are relatively easy to genotype and impute) [172]. GWAS have uncovered thousands of associations between traits and variants: as of the first quarter of 2020, the NHGRI-EBI GWAS catalog [26] contains >100,000 variant-trait associations. Individual traits or diseases may have hundreds of associated variants, most of which have a relatively small effect on the phenotype [172]; for example, a recent meta-analysis of height GWAS [311] detected > 700 associated genetic loci and the largest type 2 diabetes (T2D) GWAS to date reported > 400 independent associations [172].

While GWAS have succeeded in identifying a large number of trait-associated variants, interpretation of the associations has proved difficult. For many traits and diseases, such as T2D, most associated variants are in non-coding regions of the genome [172, 315]. The effects of these variants are often mediated by changes in gene expression, and it is frequently unclear (1) which variant(s) at the locus are the causal variant(s) (neighboring variants in high linkage disequilibrium (LD) with trait-influencing SNPs are often statistically associated with the trait as well, making it difficult to infer the causal SNPs through statistical association alone); (2) what the proximal effects of the variant(s) are; (3) which cell type(s) and conditions the causal variant(s) are having an effect in (non-coding regions of the genome are frequently functional in only a subset of cell types, developmental time points, and environmental conditions [7, 65, 184]); and (4) which gene(s) the causal variant(s) are affecting.

Because a given polygenic trait may have hundreds of associated genetic loci, targeted experimental follow-up of individual loci is laborious and costly. Therefore, a common strategy to nominate functional genetic variants at each locus and interpret their effects in different cell types is to integrate genetic variation data with genome-wide assays of molecular traits. Since it is assumed that gene expression is the vector through which many variants affect traits, these molecular traits often capture some aspect of gene regulation or gene expression. This may be gene expression itself (i.e., some measure of the ‘transcriptome’, which is the whole of the RNA produced in the cell), or some indicator of DNA gene regulatory function (often a dimension or correlate of the ‘epigenome’, which refers to the whole of the chemical marks on the DNA or on the histone proteins that form a complex with the DNA).

1.2 Chromatin accessibility as a proxy for non-coding DNA function

One layer of information that is used to decipher the function of non-coding genetic variants is chromatin accessibility data. DNA in the cell nucleus forms a complex with histone proteins, giving rise to a higher-order structure called chromatin [167]. Chromatin structure is not uniform across the genome [149, 115]. In any given cell type most of the chromatin is compacted tightly into a form called heterochromatin (‘closed’ chromatin) while a small fraction of it is in a relatively decompacted form, euchromatin (‘open’ or ‘accessible’ chromatin). Chromatin accessibility is an important factor in gene regulation [164, 280]. Gene expression is regulated through the interaction of transcription factor proteins (TFs) with regulatory elements encoded in DNA such as enhancers and promoters. For a gene to be expressed, the chromatin at the gene’s regulatory element(s) must be accessible so that the necessary TFs can interact with the DNA. Regions of accessible chromatin frequently correspond to active or primed gene regulatory elements [280], and assaying chromatin accessibility across the genome is therefore one common method of mapping likely gene regulatory elements in a cell type or tissue. Chromatin accessibility covaries with other molecular traits, including DNA methylation levels and multiple histone modifications [103, 268, 243] which themselves are indicators of regulatory function [243, 28, 48, 227].

Since trait-associated genetic loci frequently lie in non-coding regions of the genome, the causal variants at these loci are presumed to act through gene regulation and will often be in gene regulatory elements. Determining which variants lie within accessible chromatin is one method for narrowing in on the causal variants(s) at a given locus. Furthermore, because many gene regulatory elements are cell type specific (or only active under certain environmental conditions or at certain developmental timepoints) [280], this can provide information about the cell type or context in which a variant is functionally relevant. This relationship between chromatin accessibility (or other molecular marks of gene regulatory elements) and trait-associated variants has been one of the most robust findings in the GWAS era. For example, T2D associated SNPs are highly enriched in pancreatic beta cell open chromatin/enhancers [71, 79, 172, 208, 209, 224, 228, 281, 286, 287], and SNPs associated with many autoimmune disorders are enriched in immune cell open chromatin/enhancers [71, 70].

1.3 Molecular traits as mediators of genotype - phenotype relationships

Because non-coding variants act on a phenotype via intermediate effects on factors such as regulatory element function and gene expression, profiling gene expression and/or molecular correlates of regulatory element function can generate hypotheses about the precise mechanisms through which causal variants impact a phenotype. For example, if a trait-associated SNP associates with the expression of a nearby gene (such SNPs are known as expression quantitative trait loci, or eQTL), one possible hypothesis is that the effect of the SNP on the gene is the basis for the SNP-phenotype relationship. As expected, GWAS signals sometimes colocalize with eQTL [307, 104, 18, 4] or other molecular phenotype-associated variants such as chromatin quantitative trait loci (caQTL) [18, 4]. Furthermore, correlation between gene expression or chromatin accessibility and a given trait can help nominate genes or regulatory elements that might mediate genetic effects on the phenotype. Identifying associations between variants and molecular traits and between molecular traits and phenotypes can help explain variant - phenotype associations.

1.4 Non-coding variation in monogenic traits and disease

As discussed above, GWAS are used to study polygenic traits and have uncovered large numbers of phenotype-associated genetic variants, many or most of which lie in non-coding regions of the genome and have small effect sizes. Monogenic traits differ along each of these dimensions. Associated genetic variants are detected by other means, and the traits usually fall under the control of single genetic variants or mutations of large effect. Most genetic variants known to associate with monogenic diseases are coding variants [41], and as a result non-coding variation has not been emphasized in the study of monogenic traits to the extent that it has been in polygenic traits. Nevertheless, the role of non-coding variants in monogenic traits should not be overlooked, and epigenomic and transcriptomic profiling can provide valuable information in this context as well.

One example, related to later work in this thesis, can be found in the work of Weedon et al. [296]. Weedon et al. utilized whole genome sequencing of individuals with pancreatic agenesis to identify likely causal mutations, some of which clustered in a non-coding region of the genome. Epigenomic data in human embryonic stem cell-derived pancreatic endoderm cells suggested that these mutations were within a pancreatic developmental enhancer, and follow-up experiments suggested that this enhancer drives expression of the nearby *PTF1A* gene.

In other examples, profiling gene expression has clarified the impact of non-coding, disease-linked variants that alter RNA splicing [152] or RNA stability [210], and chromatin interaction data has identified pathological re-arrangements in non-coding regions of the genome [169, 168]. Transcriptomic and epigenomic profiling are therefore useful tools in understanding many monogenic disorders in addition to common polygenic traits. Data integration can help identify causal variants and resolve the path from the causal variant to the physical manifestation of the trait or disease.

1.5 Measuring genome-wide chromatin accessibility

Because chromatin accessibility is a useful indicator of regulatory activity in a cell type, investigators frequently wish to assay chromatin accessibility on a genome-wide scale. Several genome-wide assays for chromatin accessibility have been developed in the last two decades, including MNase-seq [313], DNase-seq [20], FAIRE-seq [84], and ATAC-seq [23]. ATAC-seq is the youngest of the common methods and has several practical advantages of the others, namely low input material requirements and quick turnaround time [23]. In ATAC-seq, nuclei are isolated and exposed to Tn5 transposase, an enzyme that preferentially cuts DNA where it is unprotected by histone proteins. An adapter is appended to the cleaved DNA, generating a DNA library in which the ends of the fragments represent genomic regions with accessible chromatin. Sequencing this library therefore allows one to determine which regions of the genome have accessible chromatin.

1.6 Profiling the transcriptome

Gene expression is one of the most commonly assayed molecular traits, and there are efficient, well-developed techniques to measure it. The current dominant method for measuring gene expression genome-wide is RNA-seq [160, 193, 191]. In RNA-seq, RNA is isolated from a sample and reverse transcribed to DNA, which is then sequenced. Because most RNA in the cell is rRNA [292], some form of selection (poly-A tail capture [193, 191], or negative rRNA selection [160]) is often applied to enrich the captured RNA for mRNA [318] or other RNAs of interest. Sequencing the resulting library yields a digital readout of gene expression.

1.7 Single-cell and single-nucleus assays of molecular traits

Tissues are composed of mixtures of different cell types, and even ostensibly homogeneous cell lines or isolated cell populations may contain individual cells at different stages of development or in different states. Assays that are performed on bulk tissue samples, such as standard ATAC-seq, essentially average over the constituent cell types or cell states of the input sample, obscuring the heterogeneity therein. In many contexts it is desirable to maintain and observe the per-cell heterogeneity, and as a result the development of single-cell and single-nucleus assays, which provide per-cell or per-nucleus readouts, has been an area of intense research in the last decade. The first single-cell RNA-seq (scRNA-seq) paper was published eleven years ago [276] and the technology has matured considerably since then. Single-nucleus ATAC-seq methods are younger – the first single-nucleus ATAC-seq publications appeared in 2015 [25, 49] – but are rapidly gaining adoption. Single-cell assays have also been developed to profile DNA methylation [92], histone modifications [119, 245], TF binding [119, 245] and 3D chromatin interactions [194, 229]. Recently, single-cell assays jointly profiling multiple molecular traits in individual cells have been developed as well [320, 30, 37, 44]. While these assays come with their own challenges [170], they can provide a higher-resolution view of gene expression, chromatin accessibility, and other molecular traits within a tissue or cell type of interest.

1.8 Thesis outline

In this thesis, I apply these assays – bulk and single-nucleus ATAC-seq and RNA-seq – and integrate the epigenomic and transcriptomic maps with genetic data to examine monogenic (chapter II) and polygenic (chapters IV, V) traits in human, mouse, and rat. In chapter II of my thesis, I apply ATAC-seq in concert with RNA-seq to a mouse model of human birth defects to investigate the downstream effects of a known causal genetic variant on the molecular phenotypes. In chapter III, I introduce a software package for quality control (QC) of ATAC-seq and snATAC-seq data and use it to survey public datasets as well as explore the effect of two technical variables on ATAC-seq results. In chapter IV, I utilize snATAC-seq and snRNA-seq to map the transcriptome and chromatin accessibility landscape of human and rat skeletal muscle, integrating the results with GWAS data to nominate causal variants for type II diabetes (T2D) and serum creatinine level. In chapter V, I apply bulk ATAC-seq and RNA-seq to a rat model for aerobic exercise capacity to investigate correlation between genetic variants and molecular phenotypes, as well as between molecular phenotypes and organismal phenotypes. Together, these projects demonstrate the value of epigenomic and transcriptomic data in the investigation of monogenic and polygenic traits, as well as the challenges and limitations of applying epigenomic and transcriptomic data in this context.

CHAPTER II

Genome-Wide Chromatin Accessibility and Transcriptome Profiling Show Minimal Epigenome Changes and Coordinated Transcriptional Dysregulation of Hedgehog Signaling in Danforth's Short Tail Mice

2.1 Abstract

Danforth's short tail (*Sd*) mice provide an excellent model for investigating the underlying etiology of human caudal birth defects, which affect 1 in 10,000 live births. *Sd* animals exhibit aberrant axial skeleton, urogenital and gastrointestinal development similar to human caudal malformation syndromes including urorectal septum malformation, caudal regression, vertebral-anal-cardiac-tracheo-esophageal fistula-renal-limb (VACTERL) association and persistent cloaca. Previous studies have shown that the *Sd* mutation results from an endogenous retroviral (ERV) insertion upstream of the *Ptf1a* gene resulting in its ectopic expression at E9.5. Though the genetic lesion has been determined, the resulting epigenomic and transcriptomic changes driving the phenotype have not been investigated. Here, we performed ATAC-seq experiments on isolated E9.5 tailbud tissue, which revealed minimal changes in chromatin

accessibility in *Sd*/*Sd* mutant embryos. Interestingly, chromatin changes were localized to a small interval adjacent to the *Sd* ERV insertion overlapping a known *Ptf1a* enhancer region, which is conserved in mice and humans. Furthermore, mRNA-seq experiments revealed increased transcription of *Ptf1a* target genes and, importantly, downregulation of hedgehog pathway genes. Reduced sonic hedgehog (SHH) signaling was confirmed by in situ hybridization and immunofluorescence suggesting that the *Sd* phenotype results, in part, from downregulated SHH signaling. Taken together, these data demonstrate substantial transcriptome changes in the *Sd* mouse, and indicate that the effect of the ERV insertion on *Ptf1a* expression may be mediated by increased chromatin accessibility at a conserved *Ptf1a* enhancer. We propose that human caudal dysgenesis disorders may result from dysregulation of hedgehog signaling pathways.

2.2 Introduction

The semidominant Danforth's short tail (*Sd*) mutation results in severe developmental abnormalities of the urogenital and gastrointestinal systems and the lower spine. These phenotypes overlap with caudal malformation syndromes in humans, including Currarino syndrome (OMIM #176450), urorectal septum malformation sequence, vertebral-anal-cardiac-tracheo-esophageal fistula-renal-limb anomalies (VACTERL) association (OMIM #192350) and caudal regression syndrome (OMIM #60 0145). Caudal birth defects affect 1:10,000 live births but their genetic etiology, intermediate molecular consequences to the epigenome and transcriptome and resulting pathophysiology are largely unknown [211]. The phenotypic overlap of the *Sd* mouse with multiple distinct human caudal dysgenesis disorders makes it an outstanding model to investigate the underlying etiology of human caudal dysgenesis. Heterozygous (*Sd*/+) mice have a shortened tail, sacral vertebral anomalies and frequent kidney anomalies but are viable and survive into adulthood [61]. Homozygous

(*Sd/Sd*) mice exhibit cessation of the vertebral column at the lumbar level and complete absence of the tail [61]. Urogenital defects include agenesis or hypoplasia of the kidneys with occasional formation of the bladder and urethra and persistence of the cloaca. Gastrointestinal anomalies include imperforate anus. Homozygous animals are born in Mendelian ratios but are not viable beyond 24 h [61]. *Sd/Sd* and *Sd*/+ embryos are visually distinguishable from wild type (WT) embryos at E11 by caudal truncation and hemorrhagic lesions in the tailbud [91]. Defects are apparent slightly earlier in *Sd/Sd* embryos. Prior to identification of the genetic lesion, differentiation of *Sd/Sd* and *Sd*/+ embryos could only be made by the increased phenotypic severity in *Sd/Sd* mutants during gestation [86]. In 2013 our group and others identified the *Sd* mutation as an 8528 bp insertion of an endogenous retroviral element (ERV) at a point 12,463 bp upstream of the Pancreas specific transcription factor 1a (*Ptf1a*) gene [290, 166, 262]. The ERV insertion results in ectopic expression of *Ptf1a* at E8.5 in the notochord, lateral plate mesoderm and tail mesenchyme that persists into the caudal notochord, tailbud mesenchyme, mesonephros and hindgut of E9.5 embryos [166]. The critical role of PTF1A in causing the *Sd* phenotype was demonstrated by knockout (KO) of *Ptf1a* genomic sequences followed by replacement of *Ptf1a* coding sequences, which resulted in attenuation and recapitulation of the *Sd* phenotype, respectively [262]. PTF1A is a basic helix-loop-helix (bHLH) transcription factor (TF) that is expressed in the pancreas and in neuronal progenitors in the cerebellum, hindbrain, neural tube (NT) and retina [135, 302, 198, 85]. Null mutations in *Ptf1a* in humans and mouse result in pancreatic and cerebellar agenesis [261, 136]. PTF1A interacts with an E-box protein and RBPJ to form the trimeric PTF1 complex. PTF1 binds a bipartite cognate site including a canonical E-box motif (CANNTG) and TC-box (TGGGAAA) [45]. PTF1 drives initial pancreatic organogenesis as well as the specification of GABAergic neurons in the NT [136, 118, 16]. In the pancreas, PTF1 subsequently drives differentiation of mature acinar cells [302]. Masui et al. characterized a *Ptf1a* autoregulatory enhancer region, which contains two PTF1 motifs located 14.8 and 13.5 kb upstream of *Ptf1a* [177]. Both enhancers drive *LacZ* expression in the dorsal and ventral pancreas and the NT [177]. NT-specific enhancers have been identified 12.4 kb downstream of *Ptf1a* coding sequences [182, 187]. Two long non-coding RNAs (lncRNAs), *Gm13344* and *Gm13336*, have been annotated in the genomic region surrounding *Ptf1a* in the mouse. Interestingly, the transcription start site (TSS) for *Gm13344* overlaps the region corresponding to the *Ptf1a* 13.5 kb upstream autoregulatory enhancer, while *Gm13336* overlaps *Ptf1a* coding sequences [177]. Our previous studies of the *Sd* mutation implicated ectopic *Ptf1a* expression in the *Sd* phenotype. The genome-wide epigenome and transcriptome alterations induced by the *Sd* mutation are largely uncharacterized. To address this issue, we measured chromatin accessibility using ATAC-seq and transcriptional profiles using mRNA-seq on E9.5 WT and *Sd*/*Sd* tailbuds [23, 259, 286]. We found minimal changes to the genome-wide landscape of chromatin accessibility—only one peak met our genome-wide 5% False Discovery Rate (FDR) threshold. Strikingly, the peak is adjacent to the *Sd* insertion site, which is located at the *Ptf1a* 13.5 kb upstream autoregulatory enhancer that overlaps with the *Gm13344* TSS. This peak is more open in *Sd* mice compared to WT. The mRNA-seq results confirmed upregulation of *Ptf1a* and *Gm13344* in *Sd* mutants and identified significantly (<5% FDR) reduced expression of genes within the Hedgehog signaling pathway. In situ hybridization confirmed downregulation of sonic hedgehog (*Shh*) as well as other notochordal markers. Further, dysregulation of caudal NT patterning was observed in *Sd*/*Sd* embryos, consistent with *Shh* downregulation. Finally, we observed increased apoptosis in the tailbud and caudal somites. Our findings suggest that ectopic *Ptf1a* expression, potentially driven by increased chromatin accessibility at an upstream enhancer, induces a cascade of transcriptional dysregulation of multiple genes in the Hedgehog signaling pathway. We propose that ectopic *Ptf1a* ultimately leads to downregulation of *Shh* signaling, degeneration of the notochord and increased apoptosis in the developing caudal region.

2.3 Results

2.3.1 ATAC-seq reveals localized changes in chromatin accessibility near the *Sd* insertion

TF binding and chromatin accessibility are strongly coupled. Closed chromatin and inaccessible DNA may prevent a TF from binding a target site; on the other hand, cooperative binding of TFs, or binding of certain ’pioneer factors’, may increase or maintain chromatin accessibility. Therefore, the over- or under-expression of TFs may alter chromatin accessibility [141, 43, 36, 264]. *Ptf1a* is strongly overexpressed in *Sd* mice and is a direct regulator of other TFs. The *Sd* insertion is located adjacent to a *Ptf1a* autoregulatory enhancer region. In order to quantify changes in chromatin accessibility in *Sd* mutant embryos, we performed ATAC-seq on E9.5 WT and *Sd*/*Sd* tailbuds. We analyzed four WT and four *Sd*/*Sd* ATAC-seq libraries, pooling tailbuds from two embryos for each library to ensure sufficient input material. ATAC-seq peak calling identified 26,620 reproducible peaks (defined as those peaks that appear in at least two of the eight libraries; see section 2.5). Correlation between all libraries was high with the mean Pearson's $r = 0.84$ for all pairwise comparisons (Fig. 2.1), demonstrating the reproducibility and quality of the libraries. A differential peak analysis uncovered little change in ATAC-seq signal genome wide, revealing only one significantly differential peak (unadjusted P-value 2.2×10^{-7}, adjusted P-value 5.8×10^{-3} , 5% FDR threshold for the analysis) [165]. This peak corresponds to the promoter of *Gm13344*, an lncRNA located 13.7 kb upstream of *Ptf1a* (Fig. 2.2A and B; Fig. 2.3). The peak called at the *Ptf1a* promoter region was not significantly different between WT and *Sd*/*Sd* E9.5 embryos at 5% FDR; however, the peak is skewed toward being more open in the WT (unadjusted P-value 4.5×10^{-4}, adjusted P-value 0.61 Fig. 2.2B). This suggests that while there are few genome-wide changes in chromatin accessibility in the Danforth mouse, local chromatin accessibility near *Ptf1a* and the *Sd* insertion site is altered.

2.3.2 *Gm13344* promoter is orthologous to a human pancreatic developmental enhancer

The mouse lncRNA *Gm13344* is not present in standard human genome annotations such as Gencode and RefSeq. In a previous study, expression of a transgene containing the *Sd* ERV and *Gm13344* in mice did not lead to the development of a short tail, indicating that *Gm13344* overexpression alone is unlikely to cause the Danforth phenotype [262]. LncRNAs are generally not conserved across species and may evolve from unstable transcription of regulatory elements [295, 207]. Masui et al. previously found that in the human genome, PTF1A binds an enhancer element ∼15 kb upstream of the PTF1A gene containing two PTF1A motif matches that are conserved in mouse (Fig. 2.2A, 2.3) [177]. Interestingly, the orthologous mouse *Ptf1a* binding sites occur near the *Gm13344* promoter, suggesting that the mouse *Gm13344* lncRNA promoter may be functionally comparable to the human enhancer. In order to determine whether this human enhancer is active during pancreatic development, we uniformly processed and analyzed publicly available ChIP-seq data for H3K4me1, a histone mark that colocalizes with enhancers, from developing human pancreatic cells and other control tissues (control tissues downloaded from Roadmap Epigenomics) [296, 100, 140]. Developmental pancreas H3K4me1 data show a clear H3K4me1 peak at the enhancer, consistent with the enhancer being active during human pancreatic development. (Fig. 2.2C). Notably, H3K4me1 ChIP-seq data from other tissues and time points display lower signal in the region orthologous to the mouse *Gm13344* promoter (Fig. 2.2D). Collectively, these results suggest that the *Sd*-insertion-mediated

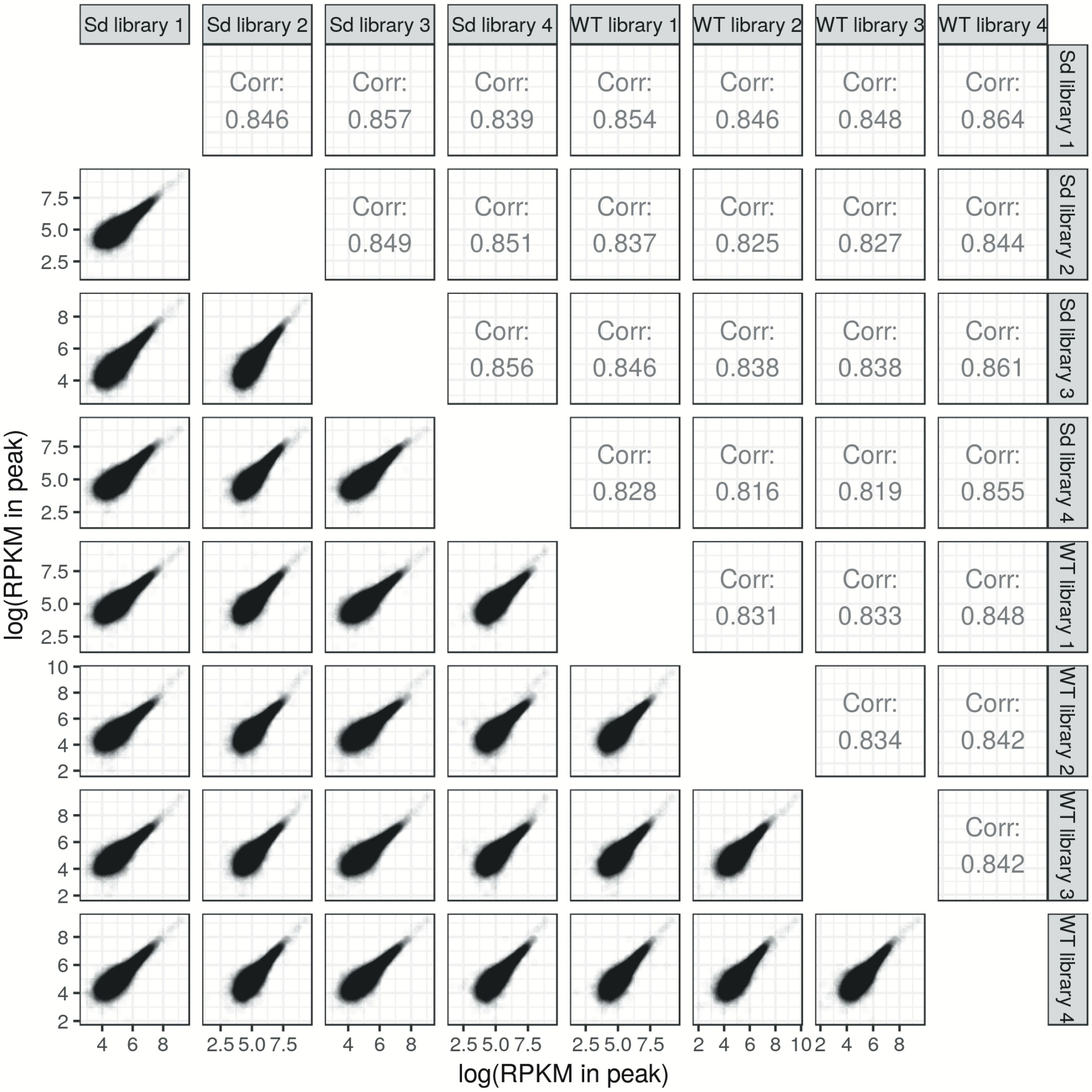


Figure 2.1 Correlation between ATAC-seq libraries. Correlation in the ATAC-seq peak signal between ATAC-seq libraries. Read counts in the 26,620 ATAC-seq peaks were converted to log(RPKM) units and the correlation between libraries calculated and plotted. Correlation values represent Pearson's r.

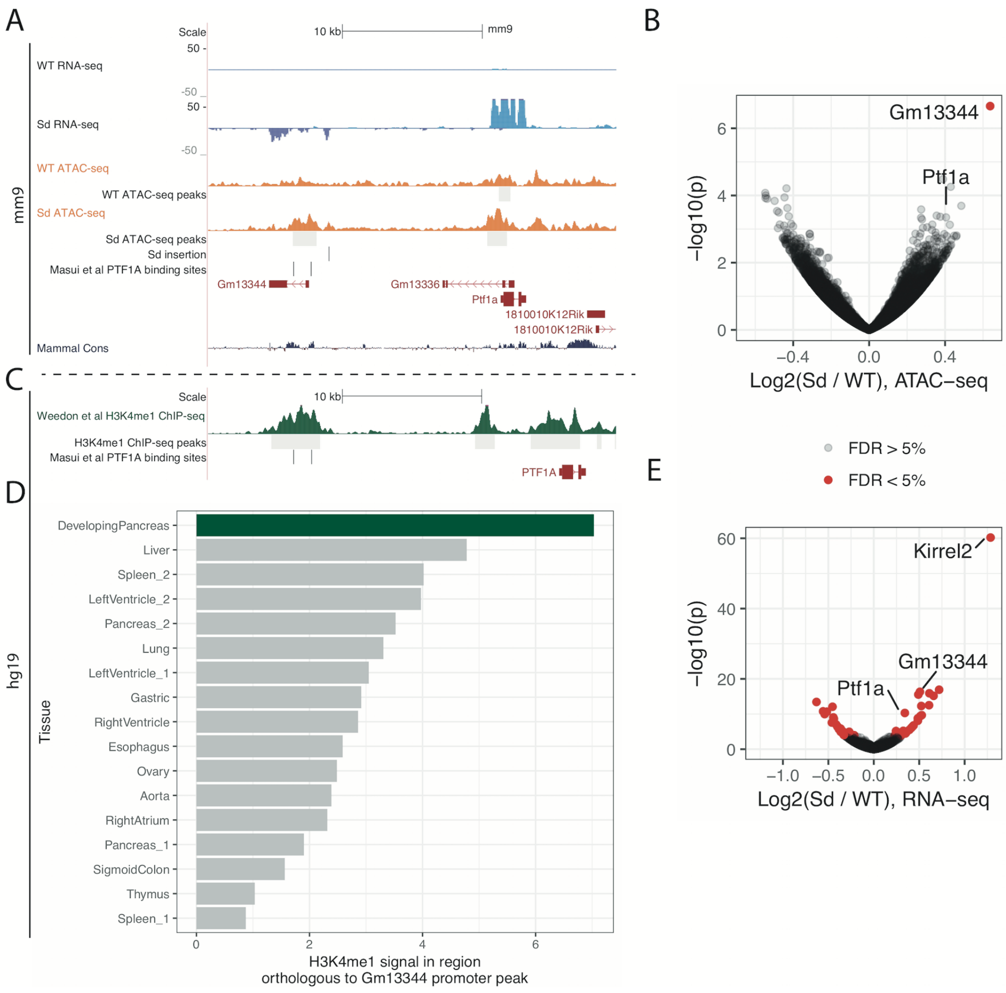


Figure 2.2 Chromatin accessibility and transcriptome changes occur at the *Ptf1a* locus of *Sd*/*Sd* mutant embryos. (A) UCSC Genome Browser [122] (`http://genome.ucsc.edu`) displaying ATAC-seq (orange) and RNA-seq signal (blue) near the Sd insertion (denoted by the vertical black line), *Ptf1a* and *Gm13344*, for one WT and one *Sd*/*Sd* library. Previously characterized PTF1A binding sites are denoted by vertical black lines near the *Gm13344* promoter and intron. (B) Volcano plot for ATAC-seq data (n = 4 libraries per genotype). The gene represented by the nearest TSS for the significantly differential peak is indicated, as is the peak representing the *Ptf1a* promoter. Red dots denote significance at FDR <5%. (C) ChIP-seq data from [296], indicating a peak above the human region orthologous to the *Gm13344* promoter. (D) The ChIP-seq signal for H3K4me1 in developing pancreas (from [296]) is greater than the signal in other tissues/at other time points. (E) Volcano plot for RNA-seq data (n = 3 libraries per genotype).

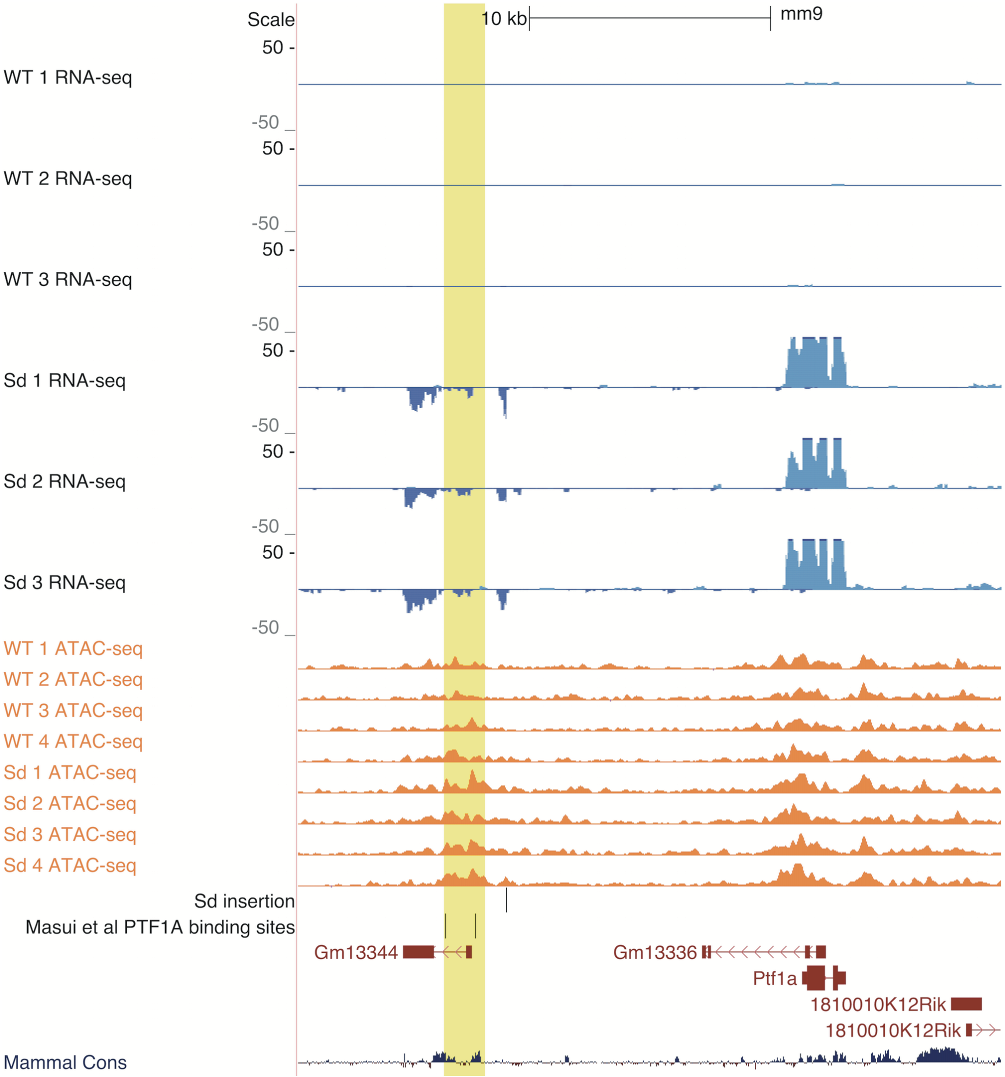


Figure 2.3 Gene expression and chromatin accessibility near the *Sd* insertion. UCSC Genome Browser [122] (`http://genome.ucsc.edu`) displaying ATAC-seq (orange) and RNA-seq signal (blue) near the *Sd* insertion, *Ptf1a*, and *Gm13344*, for all 8 ATAC-seq and all 6 RNA-seq libraries. The vertical yellow highlight represents the differential peak.

increase in chromatin accessibility at the *Gm13344* promoter acts as an enhancer for *Ptf1a* in E9.5 mice, and transcription of the lncRNA is at least partially a byproduct of this regulatory activity.

2.3.3 mRNA-seq reveals misexpression of PTF1A target genes

To determine which genes show evidence of misregulation in *Sd* mutant embryos at E9.5, we performed mRNA-seq on three WT and three *Sd*/*Sd* E9.5 tailbud libraries (samples were pooled as necessary for sufficient input material to each library; see section 2.5). We performed differential gene expression analysis and found 49 genes to be significantly differentially expressed (5% FDR; Figs. 2.2E and 2.4). Several of the differential genes discovered by mRNA-seq recapitulate previous findings. As expected, *Ptf1a* showed almost no expression in WT tailbuds but strong expression in *Sd*/*Sd* tailbuds (Figs. 2.2A and 2.4, 2.3). Similarly, *Gm13344*, the lncRNA upstream of *Ptf1a*, showed almost no expression in WT tailbuds but strong expression in the *Sd* mutant samples, consistent with previous reports (Figs. 2.2, 2.4, 2.3) [262]. Carboxypeptidase A1 (*Cpa1*), previously found to be underexpressed in E10.5 *Ptf1a* KO mice, is overexpressed in *Sd* mutant tailbuds (Fig. 2.4) [279]. We found *Foxa1* to be downregulated in *Sd* mutant samples (Fig. 2.4), consistent with the upregulation of *Foxa1* in *Ptf1a* KO mice [279]. *Foxa2* also trended toward downregulation in *Sd* mice, although the adjusted P-value did not reach statistical significance (unadjusted P-value 2.0×10^{-4} , adjusted P-value 0.066) (Fig. 2.4). Two of the most strongly upregulated genes in *Sd*/*Sd* tailbuds were *Kirrel2* and *Nphs1* (Figs. 2.2E and 2.4), which share a bidirectional promoter to which *Ptf1a* is known to bind at later time points in NT and pancreas [183]. *Aplp1*, positioned directly downstream of *Kirrel2*, is similarly upregulated in *Sd* mutant samples (Fig. 2.4). mRNA-seq also revealed upregulation of *Sox8* in *Sd* mutant tailbuds (Fig. 2.4). SOX8 is a high mobility group TF that functions redundantly with SOX9 and SOX10 in neural crest devel-

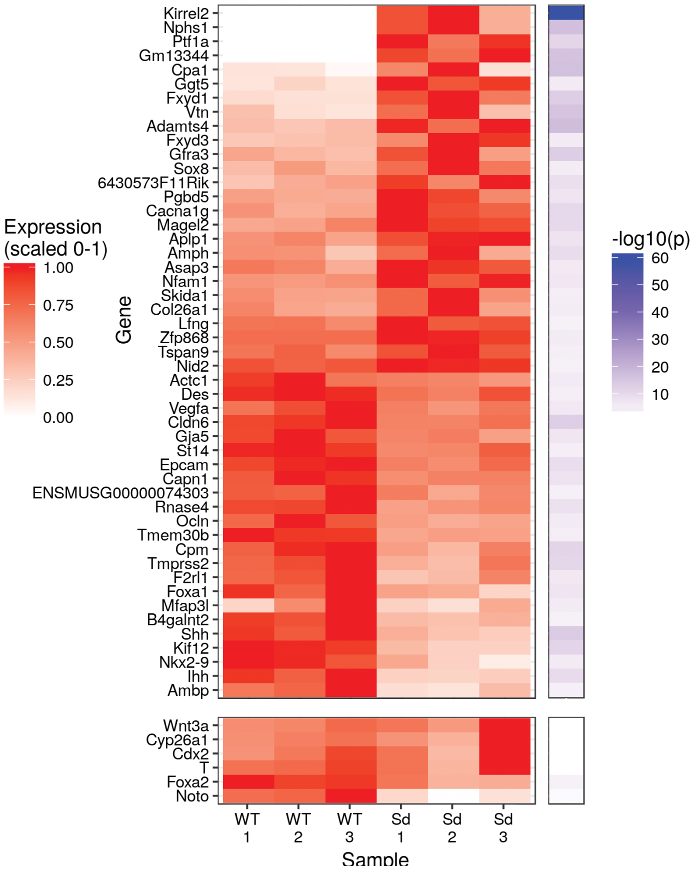


Figure 2.4 Ectopic *Ptf1a* expression drives global transcriptomic changes within Sd/Sd embryo tailbud. Heatmap displaying normalized expression and $-log10(p)$ for differentially expressed genes. Several genes that did not meet the threshold of <5% FDR but are of interest due to their known expression in the tailbud or notochord, including *Wnt3a*, *Cyp26a1*, *Cdx2*, *T*, *Foxa2* and *Noto*, are included at the bottom of the heatmap. P-values are from the differential gene expression analysis with DESeq2[165].

opment and its expression precedes that of SOX9 and SOX10 [199]. No significant differences in the caudally expressed genes *Wnt3a*, *Cyp26a1* and *Cdx2* were identified in *Sd* mutant samples in our mRNA-seq data (Fig. 2.4).

2.3.4 Whole mount in situ hybridization analysis validates mRNA-seq expression changes

To confirm expression changes in *Sd* mutant tailbuds identified by mRNA-seq analysis, we used whole mount in situ hybridization in E9.5 embryos. These studies confirmed ectopic expression of *Ptf1a* in the tailbud of E9.5 *Sd*/+ and *Sd*/*Sd* embryos (Fig. 2.5). The ectopic *Ptf1a* expression was stronger in homozygous mutants compared to *Sd*/+ embryos, and appeared highest in the presomitic mesoderm with lower levels in the notochord in the anterior thoracic region. In addition, we found increased expression of *Gm13344* in the tailbud of *Sd*/+ and *Sd*/*Sd* embryos (Fig. 2.5). We also found upregulated expression of *Kirrel2* (Fig. 2.6A–C) in *Sd*/+ and *Sd*/*Sd* mutants, as identified by mRNA-seq. *Foxa1* expression was clearly reduced in *Sd*/*Sd* embryos (Fig. 2.6D–F), in accord with our mRNA-seq data. *Foxa2* expression was also reduced in the tailbud by whole mount in situ hybridization (Fig. 2.6G–I). Although the unadjusted P-value for *Foxa2* was significant in the mRNA-seq analysis, it did not survive genome-wide multiple testing correction using 5% FDR. This could be due to the relatively small proportion of *Foxa2*-expressing cells in the tailbud [2]. We also examined expression of the T-box TF brachyury (*T*) in the tailbud of *Sd*/*Sd* embryos. *T* is expressed in notochord-derived cells and is required for notochord development, mesoderm differentiation and development of structures posterior to the forelimb level in mouse embryos [57, 257, 306]. We found that *T* expression was unaffected in the tailbud of *Sd* mutant embryos, although reduced expression was observed in the more anterior part of the notochord of *Sd*/*Sd* embryos (Fig. 2.6J–L), which correlates with our mRNA-seq data (Fig. 2.4). *Noto*, a Not family homeobox gene, is expressed in the WT mouse tailbud and is essential for proper caudal notochord function [2]. We did not detect *Noto* expression in the tailbud of *Sd*/*Sd* embryos, and expression was decreased in the *Sd*/+ embryos (Fig. 2.6M–O), which could indicate suppression of its expression, since *T* expression indicates the presence of the caudal notochord in *Sd*/*Sd* embryos at this timepoint.

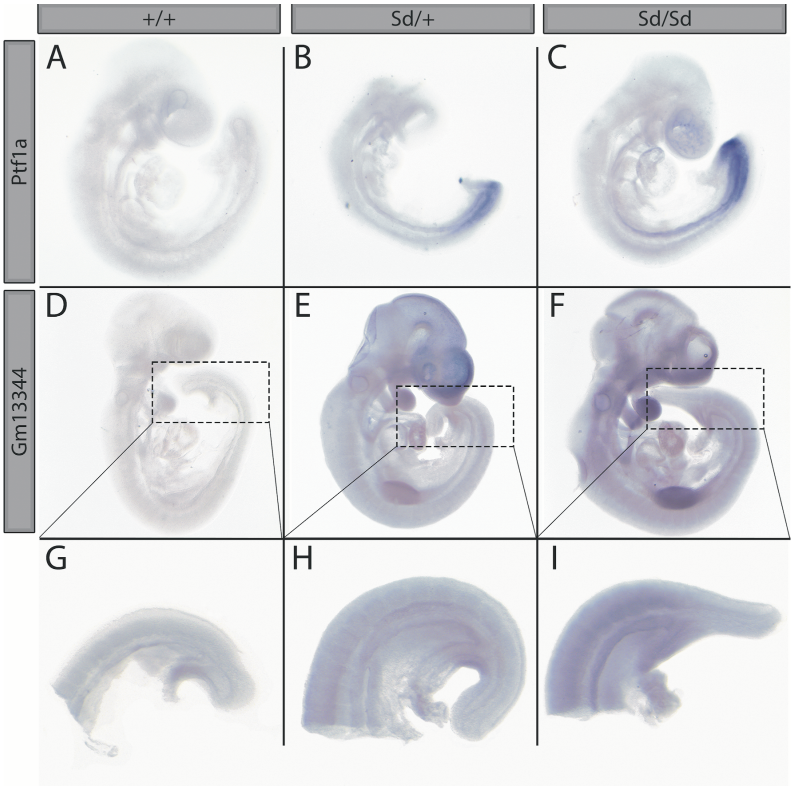


Figure 2.5 Whole mount in situ hybridization shows ectopic *Ptf1a* expression. (A-C) Whole mount in situ hybridization with *Ptf1a* antisense probe in E9.5 WT, *Sd/+*, and *Sd/Sd* embryos. (D-I) Whole mount in situ hybridization with *Gm13344* antisense probe in E9.5 embryos and tailbuds.

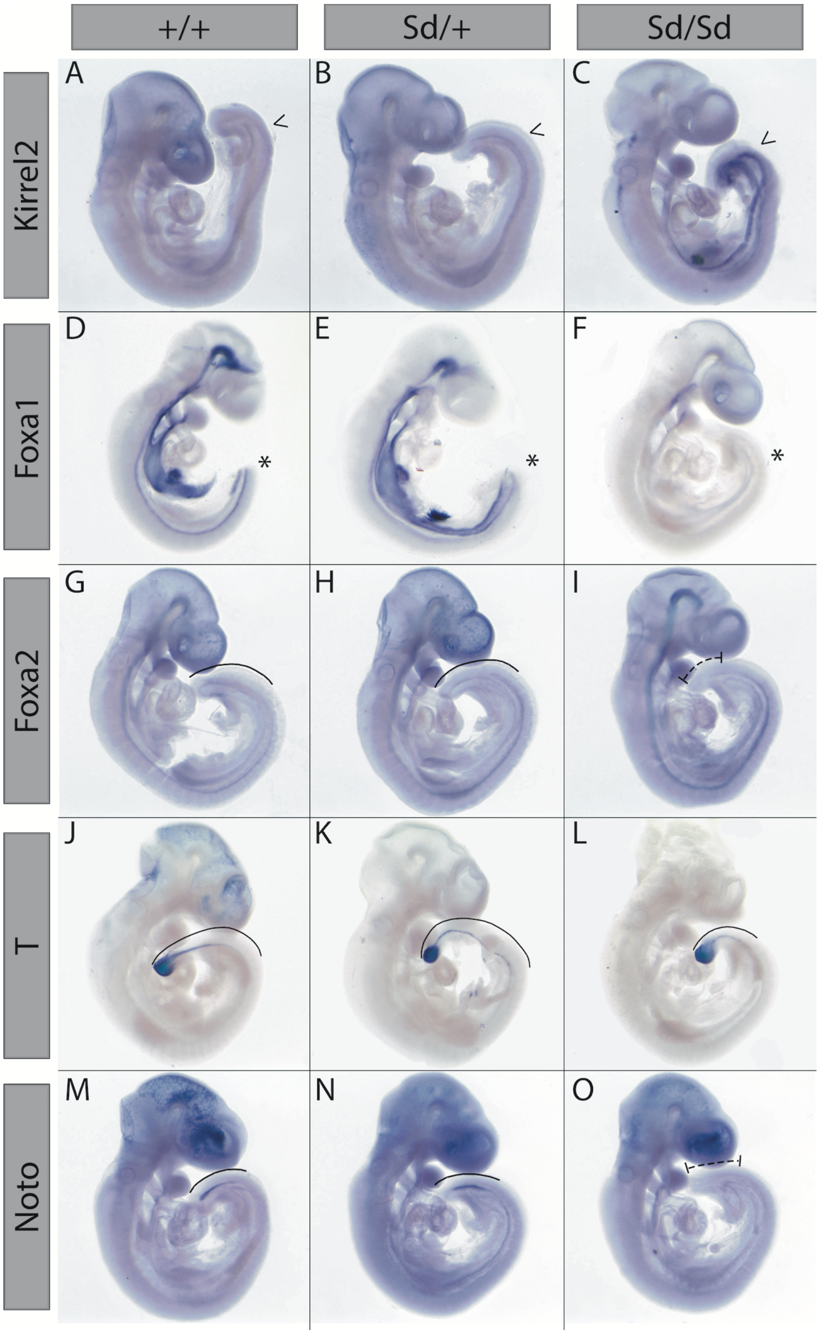


Figure 2.6 Whole mount in situ hybridization confirms transcriptomic changes in *Sd/Sd* embryos. Whole mount in situ hybridization with antisense probes for *Kirrel2* (A–C), *Foxa1* (D–F), *Foxa2* (G–I), *T* (J–L) and *Noto* (M–O) in E9.5 WT, *Sd/+* and *Sd/Sd* embryos. Arrow heads (A–C) and asterisks (D–F) indicate increased and decreased, respectively, expression patterns in the tailbud of *Sd/Sd* embryos. Solid lines indicate staining within the tailbud (G, H, J, K, L, M, N) and dashed lines (I, O) indicate extent of reduced staining in *Sd/Sd* tailbuds. $N \geq 3$ for each genotype per probe.

2.3.5 Pathway analysis shows dysregulated hedgehog signaling

To determine which developmental signaling pathways are disrupted in *Sd* mutant tailbuds, we performed a Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis using the P-values of all genes included in the differential gene expression analysis. Several pathways related to cellular adhesion and migration show significant enrichment for disrupted gene expression (Fig. 2.7A). Interestingly, the only developmental pathway to show significant enrichment for altered gene expression was the Hedgehog signaling pathway (Fig. 2.7A). *Shh* and Indian Hedgehog (*Ihh*) are two of the most strongly downregulated genes in *Sd* mutant samples (Fig. 2.4A), and in total seven genes annotated to hedgehog signaling show an unadjusted P-value for differential expression of <0.05 [*Shh* ($P = 4.43 \times 10^{-14}$), *Ihh* ($P = 6.62 \times 10^{-10}$), *Gli1* ($P = 0.018$), *Gli3* ($P = 0.019$), *Ptch1* ($P = 0.021$), *Ptch2* ($P = 0.002$), *Bmp4* ($P = 0.006$)] with changes in the direction expected for downregulation of this pathway. Whole mount in situ hybridization confirmed that *Shh* expression was absent in the tailbud of *Sd*/*Sd* embryos at E9.5, and its expression decreased in a reciprocal manner to the ectopic expression of *Ptf1a* in the tailbud (Fig. 2.7B–D). NKX2.9 is a homeobox TF expressed in close proximity to SHH in NT structures along the entire neuraxis and is regulated by SHH [205, 206]. Nkx2.9 expression was significantly reduced in *Sd*/*Sd* embryos by mRNA-seq and confirmed by whole mount in situ hybridization (Fig. 2.4A and Fig. 2.7E–G). *Gli1*, a major downstream effector of SHH signaling, was reduced by *LacZ* staining in mice from a cross between *Sd*/+ and *Gli1*-*LacZ* reporter mice (Fig. 2.7H, I) in accordance with SHH downregulation.

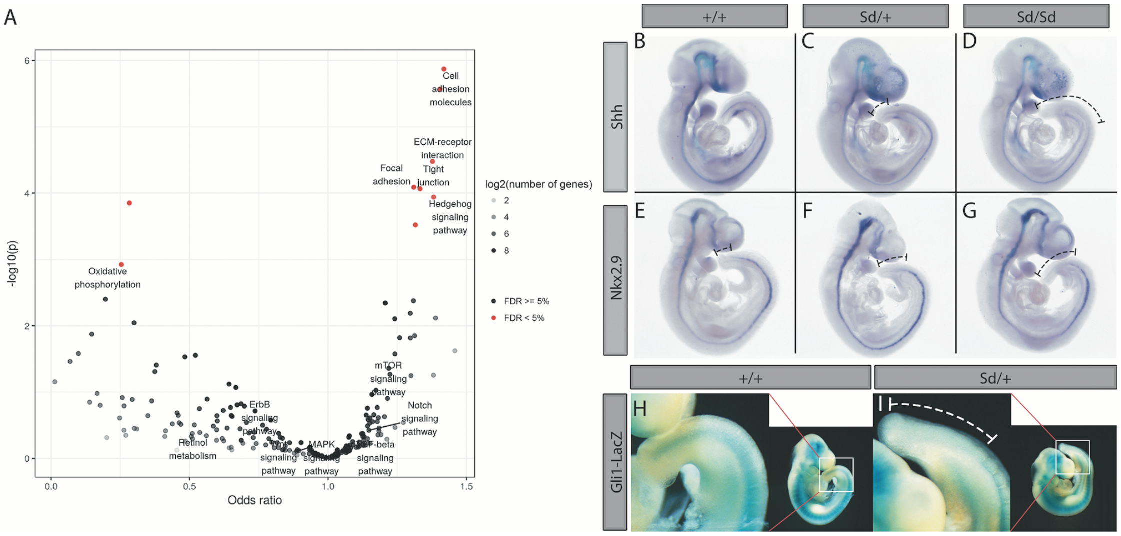


Figure 2.7 KEGG pathway analysis of *Sd/Sd* embryos identifies dysregulated SHH signaling. (A) Volcano plot for enrichment of disrupted gene expression in KEGG pathways. Notable developmental signaling pathways are labeled. Hedgehog signaling is the only developmental signaling pathway showing enrichment for differential gene expression. (B–D) Whole mount in situ hybridization with *Shh* and *Nkx2.9* (E–G) antisense probes in E9.5 WT, *Sd/+* and *Sd/Sd* embryos. Dashed lines in C, F, D and G indicate extent of reduced staining in *Sd/+* and *Sd/Sd* tailbuds, respectively. (H, I) X-gal staining of E10.5 WT and *Sd/+* embryos from a cross between *Sd/+* mice and *Gli1-lacZ* reporter mice. White dashed line in tailbud enlargement (I) shows extent of reduced *lacZ* reporter gene expression in the tailbud of *Sd/+* embryo. $N \geq 3$ for each genotype of stained embryos.

2.3.6 The absence of *Shh* results in aberrant caudal NT patterning in *Sd/Sd* embryos

We hypothesized that dysregulation of SHH signaling in developing *Sd* mutant embryos would result in aberrant NT patterning. To test this hypothesis, we conducted immunofluorescence experiments on E10.5 WT, *Sd*/+ and *Sd/Sd* mice. Embryos were sectioned in the transverse plane through the NT and notochord at the level of both the forelimb and hindlimb bud. Staining with an antibody against SHH revealed aberrant notochordal morphology in *Sd*/+ embryos and absence of the notochord in *Sd/Sd* embryos (Fig. 2.8A–F). Floorplate expression of SHH was maintained throughout *Sd*/+ embryos; however, it was lost at the hindlimb level of *Sd/Sd* embryos. Floorplate-derived SHH is induced following FOXA2 initiation in the floorplate by notochordal-derived SHH [299, 66, 109, 161, 240]. Thus, SHH loss in the floorplate is likely due to the downregulation of FOXA2 in the floorplate (Fig. 2.8G–L, E', F'). Ventralization of the NT relies on SHH signals; thus, loss of the V3 domain marker NKX2.2 in the *Sd/Sd* hindlimb NT is consistent with SHH loss (Fig. 2.8M–R, G', H') [21]. Interestingly, the more widely expressed marker NKX6.1 is expressed but ventrally constricted in *Sd/Sd* mutant embryos (Fig. 2.8S–X). Consistent with the loss of SHH signaling from the floorplate and notochord, the dorsally restricted TF PAX6 exhibits ventral expansion at the hindlimb level of *Sd/Sd* embryos (Fig. 2.8Y–D'). Reduced expression of ventral markers as well as expansion of dorsal NT markers at the hindlimb level in the NT of *Sd/Sd* embryos indicates the notochord of *Sd/Sd* mutant embryos degenerates before correct patterning of the presumptive floorplate can be established.

2.3.7 *Sd* mutant embryo exhibit increased apoptosis in the tailbud

Previous studies have revealed hemorrhaging within the regressing tailbud of E12.5 *Sd* mutant embryos; however, degeneration of the tailbud in *Sd* mutant embryos has

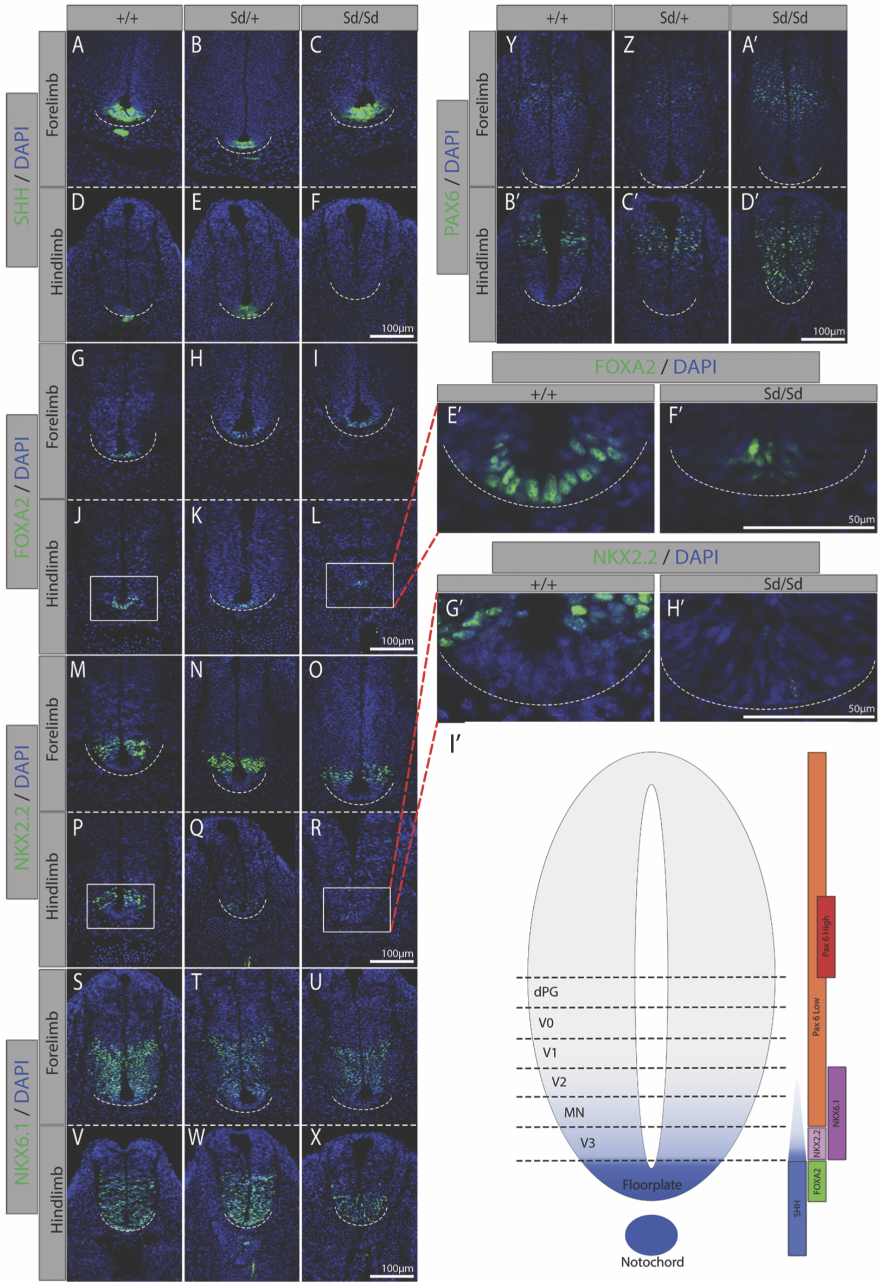


Figure 2.8 Immunofluorescence staining confirms dysregulated SHH expression in *Sd/Sd* embryos and identifies aberrant NT patterning. Immunofluorescence studies of SHH expression (green) in WT, (A, D) *Sd/+*, (B, E) and *Sd/Sd* (C, F) embryos as well as downstream targets FOXA2, (G–L) NKX2.2, (M–R) NKX6.1 (S–X) in transverse forelimb (top pane) and hindlimb (bottom pane) sections. Dorsal marker PAX6 (Y–D') indicates the competing WNT/BMP gradient. Enlarged FOXA2 (E', F') and NKX2.2 (G', H') floorplate images of WT (J, P) and Sd/Sd (L, R) hindlimb images, respectively. (I') WT expression domains of these TFs. All sections co-stained with DAPI for histological reference (blue). Dashed line indicates ventral border of NT. $N \geq 3$ for each genotype per antibody.

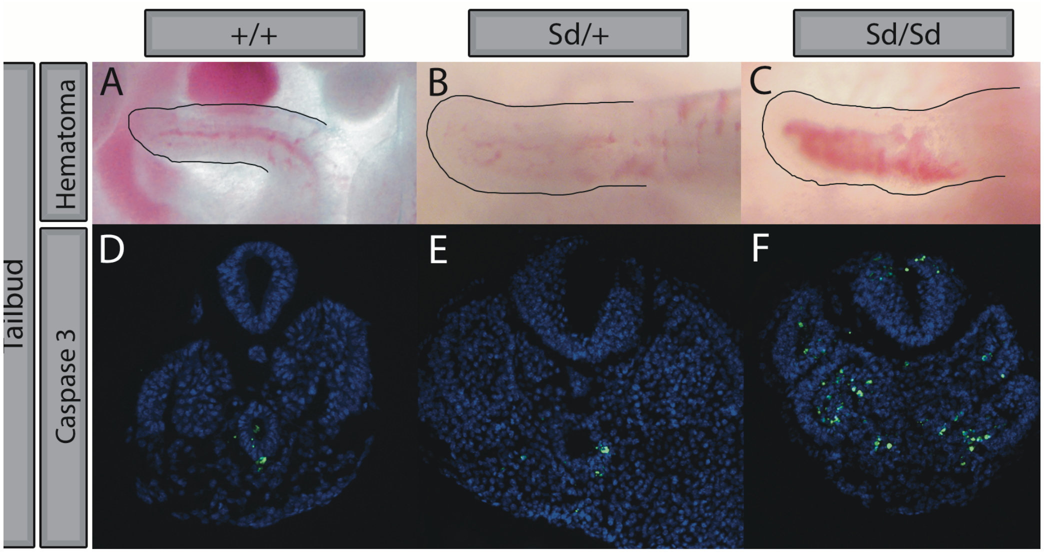


Figure 2.9 *Sd/Sd* tailbuds exhibit gross hematomas and increased apoptosis. WT (A), *Sd/+* (B) and *Sd/Sd* (C) tailbuds imaged immediately following dissection in PBS. Hematoma is evident in *Sd/Sd* embryo tailbud. Increased apoptosis was observed by immunofluorescence staining for Caspase 3 (green) in *Sd/Sd* tailbud (F) compared to WT (D) and *Sd/+* (E) tissues. All sections co-stained with DAPI (blue). $N \geq 3$ for each genotype.

not been closely investigated [86]. No visible phenotypic differences were apparent at E9.5. However, we found that hematomas are clearly evident in *Sd/Sd* embryos at E10.5, which likely give way to severe hemorrhaging as development progresses (Fig. 2.9A–C). Further investigation of the tailbud degeneration using immunofluorescence with an antibody to cleaved Caspase 3 revealed increased apoptosis in Sd/Sd tailbud tissues relative to WT and $Sd/+$ littermates (Fig. 2.9D–F). Increased apoptosis could be a result of loss of hedgehog patterning from the notochord in Sd/Sd embryos [197, 263, 294, 319, 35].

2.4 Discussion

Our group and others previously identified the *Sd* mutation as an ERV insertion upstream of *Ptf1a* that leads to its ectopic expression during caudal development, resulting in the *Sd* mutant phenotype [290, 166, 262]. The underlying mechanisms that lead to upregulation of *Ptf1a* and the downstream developmental pathways dysregulated by ectopic *Ptf1a* expression were not previously determined. In the current study, we used ATAC-seq to analyze genome-wide chromatin accessibility and mRNA-seq to gain additional insight into these underlying mechanisms. Our ATAC-seq data showed a significant change in chromatin accessibility near the *Ptf1a* genomic locus, mapping to the promoter region of *Gm13344*, an lncRNA that has been annotated in the mouse genome. This lncRNA overlaps with a previously characterized autoregulatory enhancer of *Ptf1a* [177]. According to our mRNA-seq data, *Gm13344* is transcribed and overexpressed in *Sd/Sd* tailbuds relative to WT tailbuds. Previous studies also detected overexpression of *Gm13344* in *Sd* mutant embryos, although a transgene containing the *Sd* ERV and *Gm13344* without *Ptf1a* did not recapitulate the *Sd* phenotype, while a transgene containing the *Sd* ERV and *Ptf1a* did recapitulate some, but not all, of the *Sd* phenotype [262]. We previously generated a transgene containing 31.9 kb of genomic sequence surrounding *Ptf1a* with and without the *Sd* ERV, which did not recapitulate the *Sd* phenotype [290]. This transgene contained part of the *Gm13344* sequence, including the PTF1A binding sites. The discrepancy between these transgene experiments suggests that the genomic regulatory context is critical for appropriate localization of ectopic *Ptf1a* expression and may have been strongly influenced by positional integration effects of the transgenes. Thus, the exact effects of *Gm13344* on the *Sd* phenotype are not known. We found no evidence of an annotated human lncRNA that is orthologous to the mouse *Gm13344* lncRNA. However, the promoter region is conserved in humans and maps to a region that contains high H3K4me1 signal in developing pancreas relative to other tissues, including adult pancreas [296]. This finding suggests that the enhancer properties of the *Gm13344* orthologous promoter sequence in humans are conserved. Our results suggest that the genomic sequence corresponding to the *Gm13344* promoter in mouse functions as an enhancer, consistent with previous studies, and the resulting RNA produced from *Gm13344* is a byproduct of this enhancer activity, as has been shown for other developmentally regulated genes [177, 207, 145]. In addition, we speculate that it is the combination of TFs binding to the *Sd* ERV and the *Gm13344* enhancer region that is required for the tissue and temporal specificity of ectopic *Ptf1a* expression in *Sd* mutant embryos. A corollary of this hypothesis is that ectopic *Ptf1a* feeds into this autoregulatory loop to further increase ectopic *Ptf1a* expression. One intriguing question that remains unanswered is the mechanism underlying the exquisite specificity of ectopic *Ptf1a* expression in the tailbud. We speculate that this could either be due to sequences within the ERV that direct expression of *Ptf1a* in combination with *Gm13344* enhancer sequences, or due to enhancer sequences within the surrounding genomic region, possibly within the TAD containing *Ptf1a*, that form a regulatory loop to direct this tailbud specificity. 4C-seq or Capture C experiments could shed light on this question in the future [50, 56]. Our transcriptome analysis revealed 49 genes with significant differential expression at 5% FDR. As expected, *Ptf1a* and *Gm13344* were significantly upregulated, as were known PTF1A downstream targets, including *Kirrel2*, *Nphs1* and *Cpa1*. *Kirrel2* and *Nphs1* share a bidirectional promoter, which contains known PTF1A binding sites. They were originally identified as components of the interdigitating podocyte foot processes of the glomerular filter but are also expressed in the developing nervous system [242]. Expression of both *Kirrel2* and *Nphs1* are lost in *Ptf1a* null mice, while forced expression of *Ptf1a* in the cerebral cortex induced ectopic expression of both genes [196]. Therefore, the transcriptomic changes we identified were widespread but characterized by expected changes due to upregulation of *Ptf1a* expression. It is notable that a considerable number of genes were differentially expressed while only a single region of differential chromatin accessibility was identified. Chromatin accessibility can be an important factor in regulating gene expression, but there are many other factors (such as miRNA binding, mRNA stability/degradation and TF availability) and the correlation between chromatin accessibility and gene expression is known to be poor [20]. Changes in TF binding at a promoter can often occur without changes in the ATAC-seq signal; recent work shows that the majority of TFs do not have a noticeable impact on chromatin openness [11]. As *Ptf1a* overexpression does not lead to extensive changes in chromatin accessibility, we speculate that PTF1A is one such TF. Given the above findings, it is unsurprising that many changes in gene expression occur in the absence of changes in chromatin accessibility. One of the most striking findings revealed by our mRNA-seq analysis was that the hedgehog signaling pathway is altered in *Sd* mutant embryos. Both *Shh* and *Ihh* were significantly downregulated, and KEGG pathway enrichment analysis revealed the Hedgehog pathway as significantly dysregulated compared to other known developmental signaling pathways. These findings were confirmed by in situ hybridization analysis with *Shh*, as well as a cross of *Sd*/+ mice to the *Shh* reporter strain *Gli1-LacZ*, which revealed decreased β-galactosidase expression in *Sd*/+ tailbuds. We also confirmed reduced expression of *Nkx2.9* and *Foxa2*, which are regulated by SHH along the neural axis [110]. Our mRNA-seq data did not demonstrate significant differences in expression of *Cdx2*, *T*, *Wnt3a* and *Cyp26a1*, and the KEGG pathway enrichment analysis did not identify the Wnt or Retinoic acid signaling pathways as being significantly dysregulated. Because *T* is expressed in notochord-derived cells and is required for notochord development, we performed in situ hybridization analysis for *T* expression [306, 101, 213]. While we did observe reduced *T* expression in the anterior notochord in *Sd*/*Sd* embryos, expression of *T* in the tailbud was unchanged. Although our data conflict with a previous study, our analysis of tailbud tissue and use of more specific methodology (RNA-seq and in situ hybridization compared to quantitative Reverse Transcriptase-Polymerase Chain Reaction (qRT-PCR) in RNA from whole embryos) could explain this apparent discrepancy [262]. While our use of tailbud tissue for our analyses is limited by the heterogeneous nature of the tissue, our rationale was to enrich for the entire region of ectopic *Ptf1a* expression. Although this reduces our statistical power to detect differential expression of genes expressed in only a small subset of cells within the tailbud, such as *Foxa2*, our use of additional qualitative methods to evaluate changes in gene expression, such as in situ hybridization and immunofluorescence, minimizes these concerns. We observed increased expression of *Lfng* in *Sd*/*Sd* tailbuds by mRNA-seq. *Lfng* is a downstream effector of Notch1 signaling. Although RBPJ is an integral component of the PTF1 complex, its expression was unchanged in *Sd* mutant tailbuds (unadjusted P-value = 0.56), and the Notch signaling pathway was not found to be dysregulated by KEGG pathway analysis. Since different E-box proteins are used as the third component of PTF1 complex depending on the context, it is difficult to predict which E-box proteins might be dysregulated in *Sd* mutant embryos. In addition to *T*, we also examined expression of *Noto* as a marker of notochord expression [2, 217, 180]. Not surprisingly, we found that *Noto* expression was absent in the caudal notochord of *Sd*/*Sd* embryos. It is possible that the absence of *Noto* expression is due to degeneration of notochord cells in this region. However, the persistence of *T* expression in the caudal notochord suggests that *Noto* expression is suppressed either directly or indirectly by ectopic *Ptf1a*. FOXA2 induces *Shh* expression in both notochordal and floorplate tissues, and importantly, floorplate-derived SHH is sufficient to pattern the NT [299, 66, 109, 161, 240, 8]. Immunofluorescence staining through the transverse plane of forelimb and hindlimb level NT revealed normal expression of SHH in the floorplate of *Sd*/+ embryos, which is sufficient to correctly maintain the expression domains of FOXA2, NKX2.2, NKX6.1 and PAX6 in the E10.5 NT, despite aberrant notochord morphology and reduced SHH signaling from this organizing tissue. Similarly, in *Sd/Sd* mutant embryos, normal expression of SHH from the floorplate is sufficient to pattern the NT, however, only at the forelimb level. In *Sd/Sd* embryos, SHH is completely absent from the presumptive floorplate at the hindlimb level and as a result, dorsal NT TF domains of FOXA2, NKX2.2 and NKX6.1 are reduced, and expression of the ventrally constricted TF PAX6 is expanded dorsally. Taken together, these experiments demonstrate that dysregulated SHH signaling adversely affects patterning of the NT in *Sd/Sd* embryos. If dysregulated SHH signaling drives the *Sd* phenotype, then the increased phenotypic severity of *Sd/Sd* embryos could result from the degeneration of the notochord before redundancy between the floorplate and notochord has been established in the caudal *Sd/Sd* embryo. Numerous published studies in mouse models have demonstrated that caudal and urogenital malformations resembling the malformations observed in *Sd* mutants can be secondary to disrupted SHH signaling. The *Shh* KO mouse [38] exhibits abnormal axial structures including the floorplate, ventral NT and sclerotome. Additional mouse models with disrupted SHH signaling have been demonstrated to have phenotypes of anorectal, renal and vertebral malformations that are observed in *Sd* mutant mice, resembling VACTERL association [186, 128, 127]. Runck et al. [247] studied cloacal development in *Shh* KO mice and in human patients with cloacal malformations and found striking similarities between the two. KO of *Shh* in the notochord or floorplate using either *ShhCreERT2* or *Foxa2CreERT2* mice exhibits a phenotype that is strikingly similar to *Sd* mutant mice only in the notochord KO [40]. KO of both *Foxa1* and *Foxa2* in the notochord results in severe defects in formation of the axial skeleton and dorsal–ventral patterning of the NT secondary to reduced *Shh* expression [173], also phenocopying *Sd* mutant mice. Finally, *in vivo* knockdown of *T* in the notochord results in reduced *Shh* expression in the notochord and also phenocopies the *Sd* mutant phenotype [213]. Thus, the caudal phenotype observed in *Sd* mutant mice is consistent with previous reports of disrupted SHH signaling in the notochord. Collectively, our results show that the *Sd* insertion leads to widespread transcriptional dysregulation of the hedgehog developmental signaling pathway. The effect of the *Sd* insertion may be mediated by a nearby conserved *Ptf1a* autoregulatory enhancer, which shows increased chromatin accessibility in *Sd* mutant mice. More generally, our results suggest a cascade of molecular events where a single non-coding regulatory mutation propagates to perturb a critical signaling pathway. Further, our findings suggest that regulatory mutations in hedgehog signaling pathway genes could explain some human caudal malformations without known genetic etiologies. Future studies of viable mouse models, such as the *Sd* mouse, with alterations in SHH signaling, will be critical to dissecting the timing and impact of key molecular events and will help to further our understanding of human caudal malformations.

2.5 Methods

2.5.1 Animals

All animals were housed in environmentally controlled conditions with 14 h light and 10 h dark cycles with food and water provided ad libitum. All protocols were approved by the Institutional Animal Care & Use Committee at the University of Michigan and comply with policies, standards and guidelines set by the State of Michigan and the United States Government. *Sd/+* animals were maintained on an outbred CD-1 background (Charles River Laboratories, Wilmington, MA).

2.5.2 Timed pregnancies

Matings for timed embryo isolation were set up using standard animal husbandry techniques. Noon on the day of vaginal plug observation was considered E0.5. Embryos were isolated at E9.5 - E10.5. Yolk sac DNA was isolated via the HotSHOT extraction method for genotyping. Genotyping was performed as previously described

[290]. For RNA-seq and ATAC-seq, embryo tailbuds were dissected just prior to the terminal somite, snap frozen in liquid nitrogen and stored at -80°C.

2.5.3 mRNA-seq experiments

RNA was collected from the tailbuds using a dounce homogenizer and RNAeasy mini extraction kit (Qiagen, Hilden, Germany). First, samples were placed into a dounce homogenizer with 350 ul of RLT buffer and homogenizing with five strokes. The extraction procedure was followed as specified by the kit, including the optional DNAse treatment. Samples were eluted in 35 ul of RNAse free water. Libraries were prepared at the University of Michigan DNA Sequencing Core using the Illumina TruSeq stranded mRNA-seq kit.

2.5.4 ATAC-seq experiments

All procedures were done at 4°C to minimize degradation of nuclei, and following the previously published protocol except as noted here [23]. Each tailbud was disrupted in 200 ul of NIB with 0.5% Triton-X buffer by pipetting up and down. Then the samples were spun, supernatant removed and was followed by a second wash with RSB buffer. Each pellet of nuclei was then transposed using a home-made Tn5 enzyme [259, 214]. After 30 min of incubation at 37°C, each sample was cleaned up with MinElute PCR purification kit (Qiagen, Hilden, Germany).

2.5.5 Sequencing data

Libraries were multiplexed and sequenced on an Illumina HiSeq 2500 instrument. All raw and processed data have been deposited to Gene Expression Omnibus under the accession GSE108804.

2.5.6 ATAC-seq data processing

Adapters were trimmed using cta (v. 0.1.2). Reads were aligned to mm9 using bwa mem (v. 0.7.15-r1140; flags: -M) [156]. Picard MarkDuplicates (v. 2.8.1; http://broadinstitute.github.io/picard) was used for duplicate removal (options: VALIDATION STRINGENCY = LENIENT) and samtools (v. 1.3.1) was used to filter for autosomal, properly paired and mapped read pairs with mapping quality $\geq$ 30 (samtools view -b -h -f 3 - F 4 -F 8 -F 256 -F 1024 -F 2048 -q 30) [157]. Peak calling was performed using MAC2 callpeak (v. 2.1.1.20160309; options: – nomodel –broad –shift -100 –extsize 200 –keep-dup all) [316]. Peaks were filtered against a blacklist (downloaded from http://mitra.stanford.edu/kundaje/akundaje/release/blacklists/mm9-mouse/mm9-blacklist.bed.gz). For figures displaying ATAC-seq coverage, we normalized the signal to account for differences in library size. This was done by dividing each line in the 'treat pileup' bedgraph files generated during peak calling with MACS2 by the number of tags in the treatment ('total tags in treatment' as output by MACS2) and multiplying by 10 million. The bedgraph files were then converted to bigwig format using bedGraphToBigWig (v. 4) [123]. ATAC-seq quality control was performed using ataqv (v. 1.0.0) [202]. The resulting interactive HTML quality control report is available at https://theparkerlab.med.umich.edu/data/porchard/qc/danforth_ataqv_master_peaks/. To generate this report, we used the list of peaks used for differential peak calling, the TSS file (mm9.tss.refseq.housekeeping.ortho.bed.gz in the ataqv GitHub repository and the mm9 blacklist that was used for peak filtering).

2.5.7 Differential peak calling

The list of reproducible peaks for downstream analysis was generated by taking the union of all 5% FDR broad peaks across all eight samples, and keeping the union peaks that overlapped with 5% FDR peak calls from at least two of the eight ATAC-seq libraries. We additionally tested two alternate thresholds, keeping only the union peaks that overlapped with 5% FDR peak calls from at least three or four of the eight ATAC-seq libraries, rather than from at least two; these alternate thresholds similarly resulted in only a single differential peak, the same peak identified using our selected threshold. Differential peak calling was performed across all the replicates using DESeq2 (v. 1.14.1) [165]. Read counts for each sample and each peak were collected using bedtools' coverageBed (v. 2.26.0; coverageBed -counts) [225]. The number of somites in each sample was used as a covariate (in the case that the sample consisted of pooled embryos, the number of somites for the sample was set to the mean number of somites in the embryos). Differential peaks were called at 5% FDR.

2.5.8 RNA-seq processing and differential gene expression analysis

Reads were aligned to mm9 (Gencode vM1 comprehensive gene annotation) using the STAR splice-aware aligner (v. 2.5.2b; – outSAMUnmapped Within KeepPairs) [3, 58, 59]. Aligned reads were filtered to autosomal reads with mapping quality 255 (samtools view -b -h -f 3 -F 4 -F 8 -F 256 -F 2048 -q 255). We used QoRTs (v. 1.0.7) to gather read counts per sample per gene for differential gene expression analysis [96]. Differential gene expression analysis was performed using DESeq2 (5% FDR); only genes with at least one read in at least one library were included (i.e. we excluded those genes that were completely unexpressed in all libraries). The number of somites in each sample was used as a covariate. For figures displaying RNA-seq coverage, we normalized the signal to account for differences in library size. To do this, the filtered RNA-seq BAM files were converted to wiggle format using QoRTs' bamToWiggle (with options –negativeReverseStrand –stranded –sizefactor XXX), where XXX represents the corresponding sample's size factor (from QoRTs function get.size.factors).

2.5.9 GO enrichment

Gene Ontology (GO) enrichment analysis using the P-values for all genes included in the differential gene expression analysis was performed using RNA-Enrich (with the KEGG database; code downloaded from supplemental materials of [148]. Gene sets with less than five genes or more than 500 genes were excluded from the analysis; otherwise, default parameters were used. The 'baseMean' column in the DESeq2 differential gene expression output, representing the gene-wise average read count across samples after normalizing for sequencing depth, was used as the 'avg readcount' values for RNA-Enrich.

2.5.10 ChIP-seq processing

We aligned the Weedon et al. [296] and Roadmap Epigenomics [140] ChIP-seq data using bwa aln (default parameters). All reads were trimmed to 36 bps (using fastx_trimmer from FASTXtoolkit v0.0.14; `http://hannonlab.cshl.edu/fastx_toolkit`) to prevent read length from confounding comparisons. Picard MarkDuplicates was used for duplicate removal (options: VALIDATION STRINGENCY = LENIENT) and samtools was used to filter to autosomal, properly paired and mapped read pairs with mapping quality $\geq$ 30 (samtools view -b -h -F 4 -F 256 -F 1024 -F 2048 -q 30). Peak calling was performed using MAC2 callpeak (options: –broad –keep-dup all). Peaks were filtered against hg19 blacklists (downloaded from `http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeMapability/wgEncodeDacMapabilityConsensusExcludable.bed.gz` and `http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeMapability/wgEncodeDukeMapabilityRegionsExcludable.bed.gz`). We normalized the signal for each experiment to 10M reads to prevent sequencing depth from confounding comparisons. Quality control was performed using phantompeakqualtools (v.2.0); all samples included in the analysis had relative strand cross-correlation coefficient (RSC) > 0.8 and normalized strand cross-correlation coefficient (NSC) $>$ 1.05 [144, 124]. One of the two Weedon et al. H3K4me1 samples (ERR361008) appeared as a QC outlier (NSC = 1.31, RSC = 2.67) relative to other samples (mean NSC = 1.10, mean RSC = 1.47) and was therefore excluded [296]. To determine the human sequence orthologous to the Gm13344 promoter peak, we used bnMapper with an mm9 to hg19 chain file from University of California, Santa Cruz (UCSC) (downloaded from `http://hgdownload-test.cse.ucsc.edu/goldenPath/mm9/liftOver/mm9ToHg19.over.chain.gz`) to map the *Gm13344* promoter peak onto the human genome [55]. The resulting closely spaced stretches of human orthologous sequence were then merged together (bedtools merge -d 30) in order to eliminate small gaps between them, and the signal for this region in each tissue's H3K4me1 ChIP-seq experiment was used for the comparison in (Fig. 2.2C).

2.5.11 Reproducibility of computational analyses

We have created a GitHub repository that has all the code necessary to reproduce the analyses in this work. The repository is located at `https://github.com/ParkerLab/danforth-2018`.

2.5.12 X-gal staining

X-gal experiments were conducted on $Gli1^{tm2Alj}/J$ (Jackson Laboratories, Bar Harbor, Maine, #008211) mice crossed to *Sd/+* mice. Embryos were dissected in phosphate buffered saline (PBS) and fixed with glutaraldehyde solution (1.25%, 1 M EGTA; .2%: 1 M MgCl2; 2%: 25% glutaraldehyde in PBS) and rinsed in wash buffer (.2%, 1 MgCl2 ; .4%, 5% NP-40 in sodium phosphate buffer). Embryos were incubated in X-gal stain solution containing ferric and ferrocyanide ions for 2 h. Embryos were then fixed with 4% paraformaldehyde (PFA) then rinsed and stored in PBS. ($n \geq 3$ for all conditions). Images were taken with a Leica MZ10F dissecting microscope with a Fostec EKE ACE external light source.

2.5.13 Whole mount in situ hybridization

Embryos were dissected in cold diethyl pyrocarbonate (DEPC) treated PBS, fixed in 4% PFA in DEPC treated PBS, washed in PBST, dehydrated through a graded methanol series (25, 50, 75 and 100%) and processed for in situ hybridization as described previously [120, 289]. Embryos at each developmental stage for each probe were processed together, and images were captured under identical settings for comparison of staining intensity among WT, *Sd/+* and *Sd/Sd* genotypes, n $\geq$ 3 embryos of each genotype were studied with all probe conditions. *Ptf1a* probes were generated from IMAGE clone: 8861527, (MGC:170132, GenBank BC138507.1) in PCR4-TOPO. Sense probe was generated using digoxygenin labeling kit (Roche, Indianapolis, Indiana), T3 polymerase and Not I digested plasmid. Antisense probe was generated using T7 polymerase and Spe I digested plasmid. For the *Gm13344* probe, a 511 bp fragment (Chr2:19,351,134–19,351,644) of *Gm13344* lncRNA was amplified using primers GGGTGTATCACCCAGCAATC (Forward) and AAGAGGAGGAACCCAGGTGT (Reverse) from BAC DNA RP24347 M17, and cloned into pGEMT-easy vector. The antisense digoxygenin-AP probe was generated using SP6 polymerase from the Nco I digested pGEMT clone while the sense probe using T7 polymerase from Nde I digested clone. Probes for *Shh* [95, 62, 68], *Noto* [180], *Kirrel2/Neph3* [291], *Nkx2.9* [205], *Foxa1*, *Foxa2* [188, 112] and *T* [305] were previously described.

2.5.14 Immunofluorescence

Embryos were dissected into fresh, ice cold 4% PFA (Fluka: 76240) in PBS for 30 min then cryoprotected overnight in solution of 10% sucrose and 2 mM MgCl2 and cut immediately rostral to the forelimb and hindlimb buds. The tissues were embedded in optimum cutting temperature (OCT) compound and stored at -80 °C. Frozen sections were taken through the transverse plane and stored at -80 °C. Tissue sections were blocked with PBS with 3% Bovine Serum Albumin (BSA), 1% Goat Serum, 0.1% Triton X-100, brought to a final pH of 7.4 for 1 h at room temperature in a humidified container. Blocking solution was replaced with primary antibody diluted in blocking buffer. Primary antibodies were purchased from the Developmental Studies Hydronima Bank [SHH (5E1, AB 2188307), FOXA2 (4C7, AB 2278498), NKX2.2 (74.5A5, AB 2314952), NKX6.1 (F55A10, AB 532378), PAX6 (PAX6, AB 528427)]. Tissue sections from n $\geq$ 3 embryos per genotype were incubated with each primary antibody overnight at 4°C at the following concentrations: 1:20 for SHH, FOXA2, NKX2.2, NKX6.1 and 1:2500 for PAX6. Sections were stained with DAPI (Kirkegaard & Perry Laboratories, Gaithersburg, Maryland: 71-03-00), incubated for 1 h with Alexa fluor 488 goat anti-mouse secondary antibody (A11001), then coverslipped with aqueous mounting medium (Thermo Scientific, Waltham, Massachusetts: Immu-Mount: 990402). Fluorescent imaging was conducted on a Leitz DMRB microscope, with a Leica EL6000 fluorescent lamp.

2.5.15 Caspase 3 immunofluorescence

Embryos were processed for Paraffin embedding with a Tissue Tek VIP (Miles Scientific, Newark, Delaware) and embedded in paraffin blocks. Blocks were sectioned with a Spencer Microtome (American Optical, Bethlehem, Pennsylvania) at a thickness of 7 µm and stored at room temperature. Paraffin was removed with Xylene followed by EtOH and PBS for rehydration. Antigen retrieval was achieved by citrate boiling and slides were then blocked with suppressor solution (5% goat serum, 3% BSA, 0.5% Tween 20) for 20 min. Sections (n $>$ 3 per genotype) were incubated with Caspase 3 primary antibody (Cleaved Caspase-3 anti-rabbit, D175: 9661S Cell Signaling Technology) overnight at 4°C at a concentration of 1:400. Sections were stained with DAPI (Kirkegaard & Perry Laboratories: 71-03-00), incubated for 1 h with Alexa fluor 488 goat anti-rabbit secondary antibody (A11008), rinsed with PBS, then coverslips were applied with aqueous mounting medium (Thermo Scientific, Waltham, Massachusetts: Immu-Mount: 990402). Fluorescent imaging was conducted on a Leitz DMRB microscope, with a Leica EL6000 fluorescent lamp.

2.6 Acknowledgements

The authors thank Drs Sally Camper and Miriam Meisler for critical review of this manuscript. We acknowledge Dr Andreas Kispert for the *Kirrel2*/*Neph3* probe, Dr Hans-Henning Arnold for the *Nkx2.9* probe, Dr Andreas Schedl for the *Nephrin*/NPHS1 probe, Dr Klaus Kaestner for *Foxa1* and *Foxa2* probes and Dr Cheryle Seguin for the *Noto* probe. We thank the University of Michigan DNA Sequencing Core for assistance with sequencing.

This project was funded by March of Dimes (6-FY14-403 to C.E.K.); the American Diabetes Association Pathway to Stop Diabetes (1-14-INI-07 to S.C.J.P.); and the National Institutes of Health (T32 HG00040 to P.O.).

2.7 My contributions

This project is published in [201]; as indicated by the author list for that paper, this was a joint project I undertook with James S. White, Dr. Peedikayil E. Thomas, Anna Mychalowych, Anya Kiseleva, John Hensley, Prof Benjamin Allen, Prof Stephen C.J. Parker, and Prof Catherine E. Keegan. The diverse set of analyses contained within this work was only possible thanks to the efforts of that diverse team. I performed the bioinformatic analyses, including analysis of the ATAC-seq and RNA-seq data produced for the manuscript as well as analysis of public datasets such as the public ChIP-seq data. All authors wrote the manuscript; my writing contribution focused on the bioinformatic analyses and discussion of their implications.

CHAPTER III

Quantification, Dynamic Visualization, and Validation of Bias in ATAC-seq Data with Ataqv

3.1 Abstract

The assay for transposase-accessible chromatin using sequencing (ATAC-seq) has become the preferred method for mapping chromatin accessibility, due to its time and input material efficiency. However, it can be difficult to evaluate data quality and identify sources of technical bias across samples. Here, we uniformly analyze 2,009 public ATAC-seq datasets and present ataqv, a computational toolkit for efficiently measuring, visualizing, and comparing quality control (QC) results. We observed a ten-fold range across key QC metrics in the public datasets. We next performed benchmark ATAC-seq experiments and statistical modeling to show that technical variation in the ratio of Tn5 transposase to nuclei and sequencing flowcell density induces systematic bias, substantially changing the enrichment of reads across diverse functional genomic annotations including promoters, enhancers, and transcription factor bound regions, with the notable exception of CTCF. We show that key QC metrics can adjust for these technical biases and conclude that the ataqv tool and associated benchmark datasets will help increase the reproducibility and rigor of ATAC-seq conclusions.

3.2 Introduction

The assay for transposase-accessible chromatin using sequencing (ATAC-seq) is the current preferred method for mapping chromatin accessibility due to its simplicity, speed, and low input material requirements [23]. In ATAC-seq, intact nuclei are exposed to Tn5 transposase, which preferentially cuts protein-free unprotected DNA to ligate sequencing adapters to the cleaved ends. After sequencing, the reads are aligned to a reference genome and peak calling is performed to determine the regions of the genome enriched for transposase-accessible DNA. This information can be used to inform the prediction of active regulatory regions [23], nucleosome positioning [254], and transcription factor binding [223, 256].

The number of publicly-available ATAC-seq datasets is rapidly growing, but the quality of these datasets can vary widely. ATAC-seq libraries may differ in PCR amplification bias, fragment length distribution, transcription start site (TSS) enrichment, nuclei prep quality, proportion of mitochondrial reads, and other variables [14]. ATAC-seq involves a number of experimental and computational steps which may introduce such heterogeneity. Some of these confounders are shared with many other high-throughput sequencing-based assays (e.g., PCR amplification bias), while others are more ATAC-seq specific (e.g., potentially high proportions of mitochondrial reads and variable nuclei prep quality). Identifying these confounders and adjusting for them in downstream analyses is an important part of reproducible and rigorous ATAC-seq analyses.

Few computational quality control (QC) tools exist for ATAC-seq, and each of the existing tools have notable limitations. The ENCODE ATAC-seq processing pipeline includes a script (ATAqC; `https://github.com/kundajelab/ataqc`) that produces a QC report, but this script is difficult to utilize as a standalone tool. Considerable effort is required to integrate it into a custom pipeline as one must install a complete conda environment, and it supports only the human and mouse reference genomes. It produces one report per sample (rather than a unified report for multiple samples), complicating cross-sample comparisons. A second tool, ATACseqQC [203], exists as an R Bioconductor package. This package provides R functions for QC of BAM files and preprocessing for common downstream analyses. Because it provides functions rather than generating a single report, it is a flexible framework but places additional work on the end user and renders the package inaccessible to those unfamiliar with R. Like ATAqC, it generates separate plots for each BAM file, making it less practical for cross-sample comparisons. Alfred [233] is a third tool with ATAC-seq QC functionality. It is run on the command line and a web server is available for visualizing the results. It is quick to set up and run, but does not have an option to visualize several libraries simultaneously, and can handle only three read groups per BAM file in the case that the user wishes read groups to be analyzed separately. Several additional software packages built to assist in ATAC-seq data processing and analysis exist, and these each include QC steps; however, they are not meant to provide comprehensive QC on BAM files. These packages include ATAC-pipe [322], which supports only two reference genomes (hg19 and mm9) and does not perform QC on a user-provided BAM file (the primary QC function accepts raw fastq files, tying read mapping and QC together); and esATAC [298], which similarly provides few read mapping statistics when starting from a BAM file (rather than a fastq file) and produces individual QC plots (e.g., fragment length distribution and TSS coverage) for each sample/replicate, complicating cross-sample comparisons.

In order to address these shortcomings, and to facilitate the unified analysis of thousands of ATAC-seq datasets, we developed a new ATAC-seq QC and visualization software package, ataqv (Fig. 3.1). Ataqv overcomes the primary limitations of existing packages. It eases cross-sample and cross-experiment comparisons, can easily be integrated into existing data processing pipelines, and produces interactive reports that are easy to share. We apply ataqv to thousands of publicly-available libraries and observed a broad range of results across diverse QC metrics. We therefore carefully constructed Tn5 dosage experiments to explore the influence of technical variation on ATAC-seq profiles and find that experimental conditions that influence the ATAC-seq fragment length distribution, such as sequencing lane cluster density and Tn5:nuclei ratio, robustly skew QC metrics and alter the biological interpretation of ATAC-seq results. QC reports and metrics from the ataqv package can help identify these technical biases and adjust for them in downstream analyses.

3.3 Results

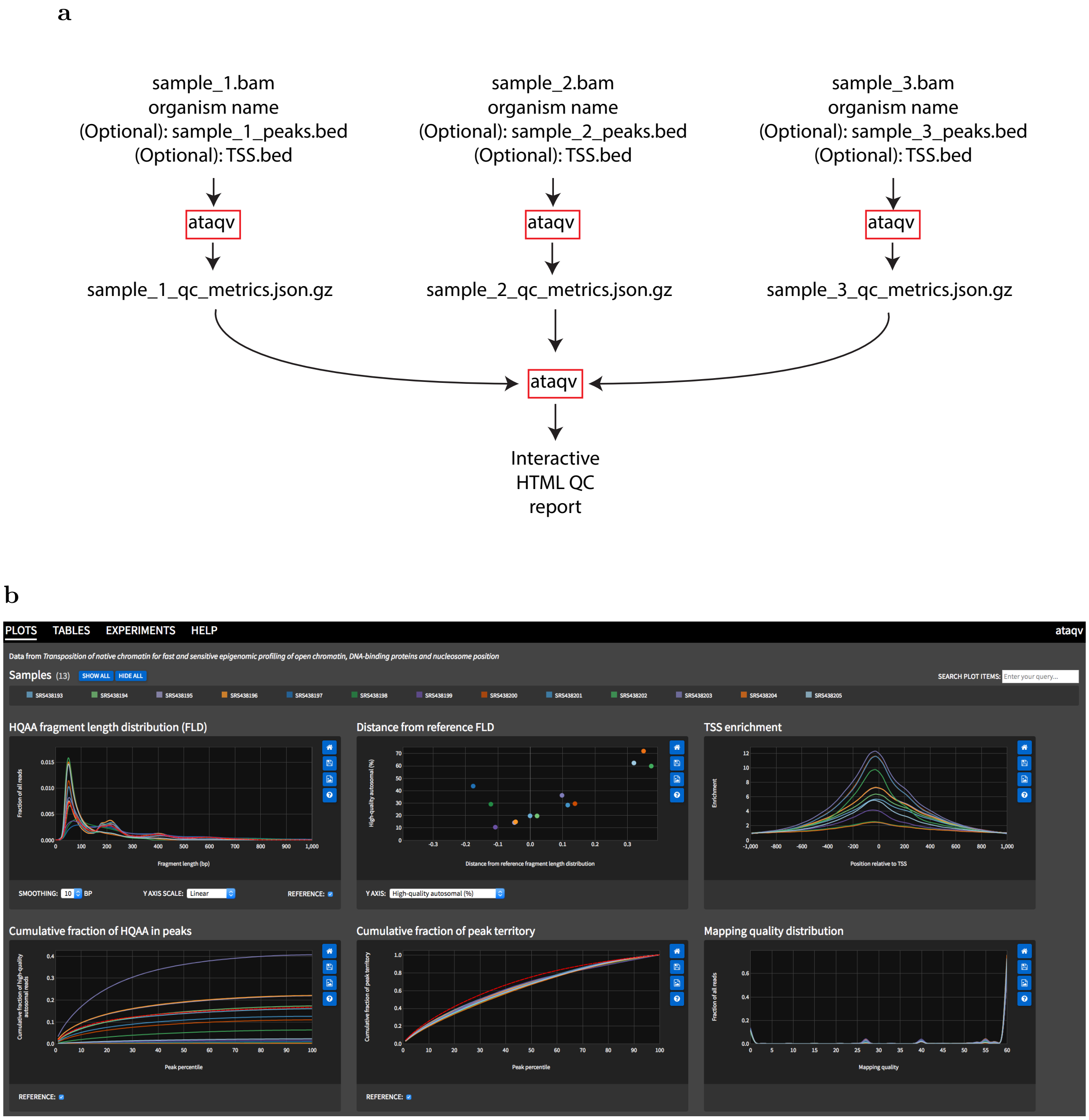
3.3.1 Ataqv is a modular and accessible tool for ATAC-seq quality control and visualization

Ataqv allows quick visualization and comparison of 35 metrics and potential confounders across samples (Table 3.1; Fig. 3.1). It produces both machine-readable (JSON format) metrics and an interactive HTML report (Fig. 3.1b) that is accessible to experimental scientists and easy to share. It is simple to integrate into existing ATAC-seq pipelines and can handle thousands of samples, an important consideration as single-cell analyses come of age and sample sizes grow. The only inputs are a BAM file of aligned reads, an optional BED file of peaks, and the name of the organism to which they were aligned (Fig. 3.1a). Human, mouse, rat, worm, fly, and yeast reference metadata is built in; metadata for other organisms (autosomal and mitochondrial chromosome names) can be easily supplied. If desired, metrics can be calculated separately for each read group in a BAM file (facilitating the processing of BAM files that may contain many libraries, as is often the case for single-cell data). A demonstration of the interactive ataqv HTML report is at `https://parkerlab.github.io/ataqv/demo/` and the ataqv source code is freely available under the GPL3 license at `https://github.com/ParkerLab/ataqv/`.

Metric	**Abbreviation in Figure 3.2**
Fragment length distribution	NA
% reads that are high-quality and autosomal	percent_hqaa
% of reads properly paired and mapped	percent_properly_paired_and_mapped
% of reads that aligned to autosomes that were duplicates	percent_autosomal_duplicate
Short-to-mononucleosomal-ratio (# fragments 50 - 100 bps long / # 150-200 bps long)	short_mononucleosomal_ratio
TSS enrichment	tss_enrichment
Duplicate fraction in peaks (fraction of properly paired reads that map within peaks and are duplicates)	duplicate_fraction_in_peaks
Duplicate fraction outside of peaks (fraction of properly paired reads that map outside of peaks and are duplicates)	duplicate_fraction_not_in_peaks
Peak duplicate ratio (’duplicate fraction outside of peaks’ / ’duplicate fraction in peaks’)	peak_duplicate_ratio
Cumulative fraction of high-quality autosomal reads in peaks	hqaa_overlapping_peaks_percent
Cumulative fraction of the genome that falls within peaks	total_peak_territory
Distribution of mapping qualities	N/A
Number of total reads	total_reads
% of alignments marked secondary	percent_secondary

% of alignments marked supplementary	percent_supplementary
% of alignments marked as duplicates	percent_duplicate
Mean mapping quality	mean_mapq
Median mapping quality	median_mapq
% of reads unmapped	percent_unmapped
% of reads with an unmapped mate	percent_unmapped_mate
% of QC-fail reads	percent_qcfailed
% of unpaired reads	percent_unpaired
% of reads with mapping quality 0	percent_mapq_0
% paired + mapped but in RF orientation	percent_rf
% paired + mapped but in FF orientation	percent_ff
% paired + mapped but in RR orientation	percent_rr
% paired + mapped but on separate chromosomes	percent_mate_separate_chromosome
% paired + mapped but too far from mate	percent_mate_too_distant
% paired + mapped but not properly	percent_improperly_paired
% reads aligned to autosomes	percent_autosomal
% reads aligned to mitochondria	percent_mitochondrial
% reads aligned to mitochondria that were duplicate	percent_mitochondrial_duplicate
Number of peaks called	total_peaks
Fragment length distribution distance	fragment_length_distance
Max fraction of reads from a single autosome	max_fraction_reads_from_single_autosome

Table 3.1: List of ATAC-seq metrics displayed in interactive ataqv HTML report



Data from *Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position*

SMOOTHING: 10 BP   Y AXIS SCALE: Linear   REFERENCE: ☑

Y AXIS: High-quality autosomal (%)

REFERENCE: ☑

REFERENCE: ☑

Figure 3.1 Ataqv workflow. Aligned reads (BAM format), the organism name, and optionally a file of peak calls and TSS annotations are passed to ataqv, which generates a JSON-formatted file of quality control metrics. JSON files for multiple BAM files can be passed back to ataqv, which then creates an interactive HTML report displaying the samples jointly.

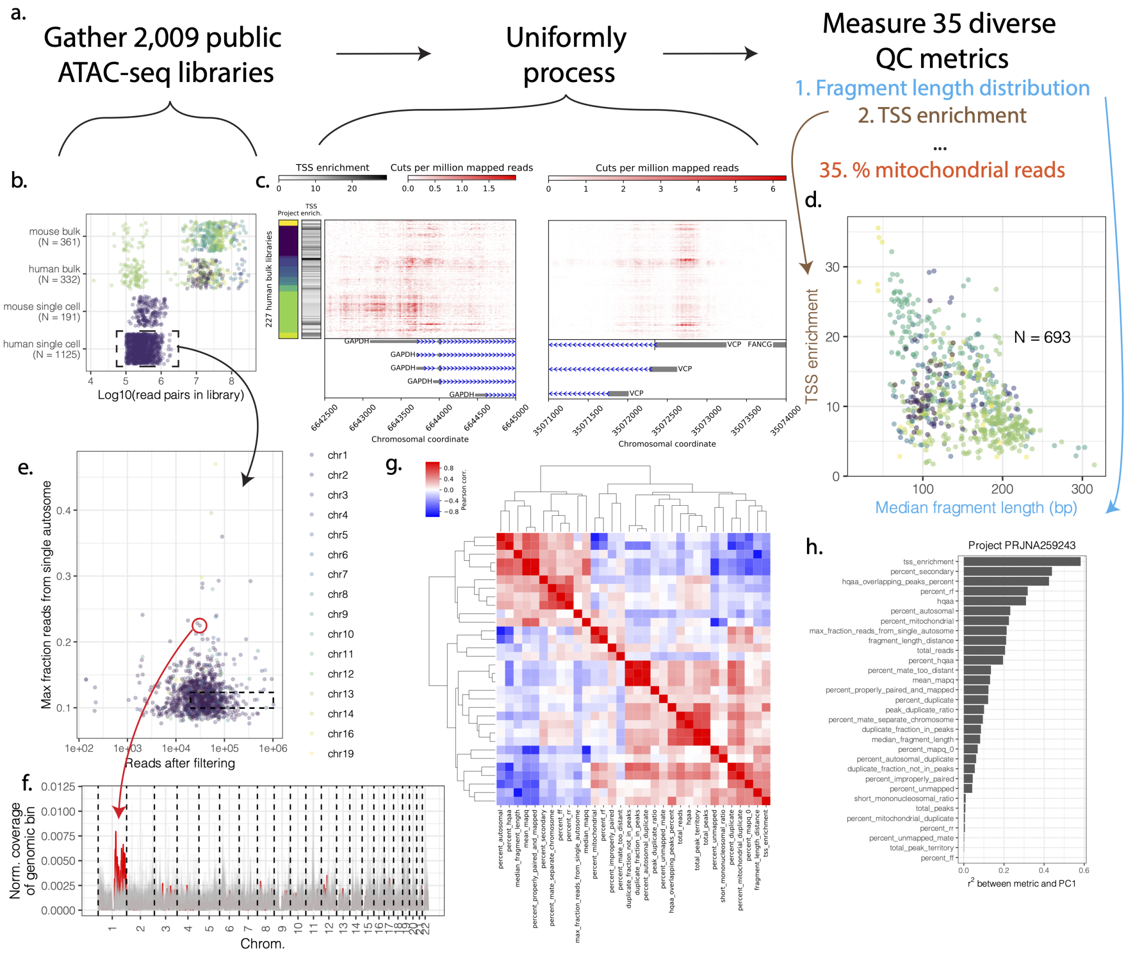


Figure 3.2 Survey of public ATAC-seq data (a) 2,009 public ATAC-seq libraries representing 23.4 billion read pairs were downloaded and uniformly processed. (b) Number of libraries and total read pairs per species and project (colors represent different projects). (c) ATAC-seq signal at promoters of two housekeeping genes (*GAPDH* and *VCP*) across human bulk libraries with at least 5M reads post-filtering. Colors along the y-axis represent project. (d) TSS enrichment and median fragment length for the 693 processed bulk (not single cell) datasets. (e) Maximum fraction of autosomal reads derived from a single autosome for public human single-cell ATAC-seq data. (f) Normalized read coverage in 2Mb windows (with 1Mb steps between them) across chromosomes for the outlier circled in red from (e) and for a set of 90 non-outlier cells from the same cell type (GM12878; all lying within the dotted box in (e)). The outlier's read coverage is represented by the red line; non-outliers are shown in gray. One arm of chromosome 1 shows abnormally high coverage in the outlier cell. (g). Correlation between ataqv metrics across public bulk ATAC-seq datasets. Metric abbreviations are listed in Table 3.1 (h). Correlation between PC1 and ataqv metrics in project PRJNA259243.

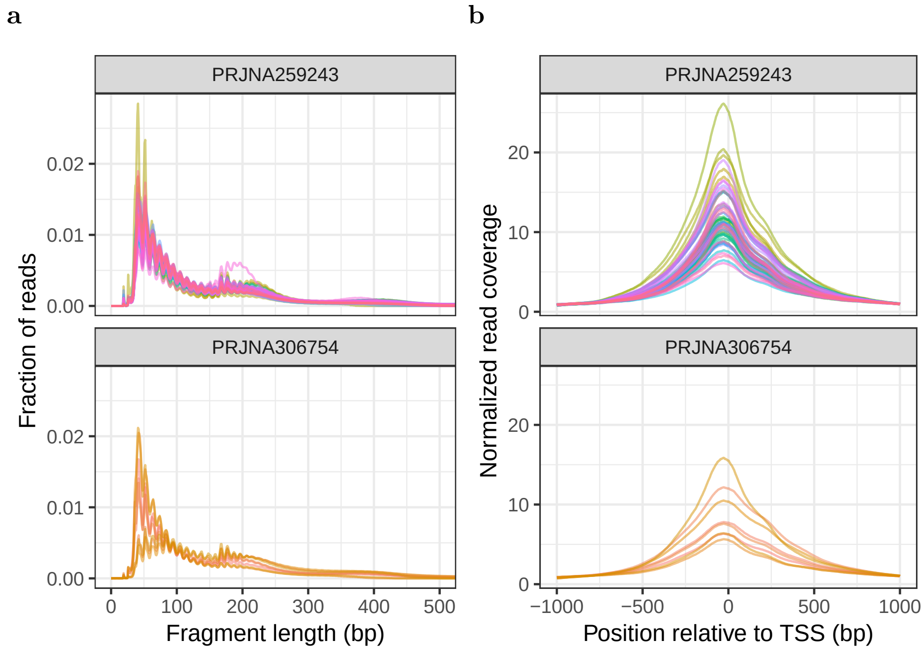


Figure 3.3 Intrastudy heterogeneity in ATAC-seq data. (A) Fragment length distributions and (B) TSS enrichment for two public datasets (facet header labels).

To demonstrate the utility of ataqv and assess the heterogeneity of publicly-available datasets, we downloaded and uniformly processed 2,009 human and mouse ATAC-seq libraries (Fig. 3.2a-c). The fragment length distributions (FLDs) and TSS enrichment for these libraries display over ten-fold variability (Fig. 3.2d), and considerable heterogeneity exists even between libraries from the same study (Fig. 3.2c,d, 3.3). Links to the interactive ataqv sessions for these uniformly processed data sets are available in section 3.5. These sessions make clear the heterogeneity in public ATAC-seq data and may be helpful as a point of reference to compare new ATAC-seq datasets.

QC of single-cell ATAC-seq (scATAC-seq) data is especially critical to ensure meaningful and reproducible results, as a portion of the sequencing reads produced in scATAC-seq experiments may be derived from background DNA released by non-viable cells. Per-cell scATAC-seq fragment counts, sometimes in combination with TSS enrichment, is commonly used to filter the data to those reads derived from high-quality cells [228, 252]. Ataqv introduces another metric that we believe will be useful in filtering single-cell data in particular: the maximum fraction of autosomal sequencing reads derived from a single autosome. Most scATAC-seq data is produced using microfluidics platforms, and in such systems free DNA from dead or dying cells may end up being transposed and barcoded, perhaps in close proximity to (and with the same barcodes as) healthy cells. As a result some cellular barcodes that appear to represent healthy cells and could pass common QC thresholds may contain reads from this free DNA. If large, free-floating chromosomal segments are represented in such barcodes, the observed distribution of reads across chromosomes would not match the expected distribution. To demonstrate this, we examined the maximum fraction of autosomal sequencing reads derived from a single chromosome for all cells in public scATAC-seq data (Fig. 3.2e). This metric illuminated extreme chromosomal read imbalance in some of the cells. Plotting the read coverage in genomic bins across each chromosome for some of these cells frequently showed that large contiguous segments of the chromosomes have increased coverage relative to the rest of the chromosomes (Fig. 3.2f), consistent with a scenario in which broken chromosome(s) derived from another cell received the same barcode as the reads from a potentially healthy cell. Such cases should be filtered out during QC of scATAC-seq data. We additionally examined this metric in the public bulk ATAC-seq libraries, where outliers may reflect abnormal karyotypes (Fig. 3.4). Consistent with this notion, the outliers we observed (e.g., K562 and mESCs) tended to be cell lines with known abnormal karyotypes [195, 234, 271].

To explore the relationship between QC metrics, we calculated the correlation between all QC metric pairs across the public bulk libraries analyzed (Fig. 3.2g). We find that TSS enrichment positively correlates with percentage of reads in peaks, and negatively correlates with median fragment length. Read count positively correlates with the number of peaks called, likely because greater read count increases the

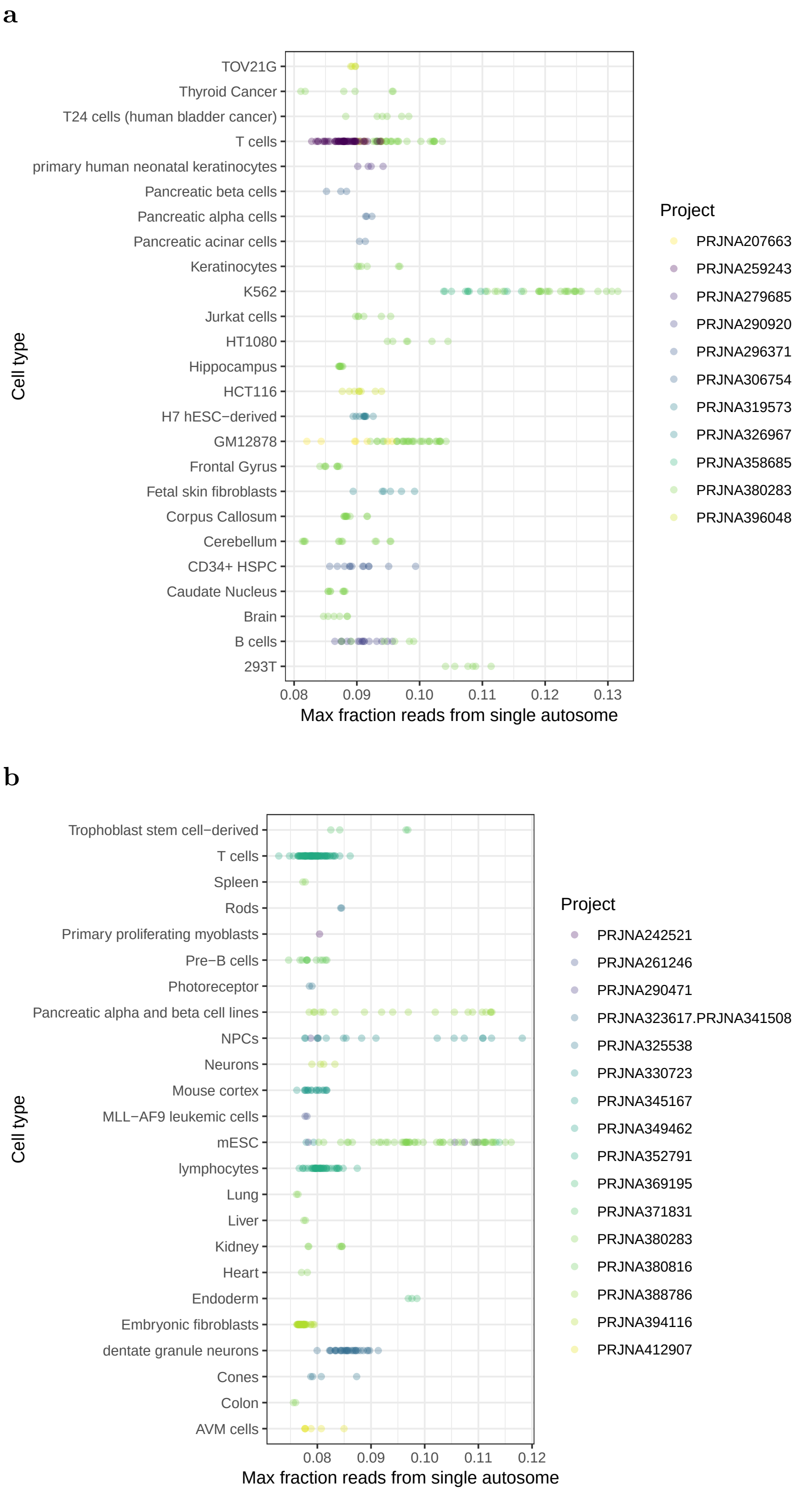


Figure 3.4 Maximum fraction of autosomal reads from a single chromosome in public bulk ATAC-seq data. Maximum fraction of autosomal reads from a single chromosome in public bulk ATAC-seq data from (A) human and (B) mouse. Each point represents one library. Outliers tend to be cell lines with known abnormal karyotypes.

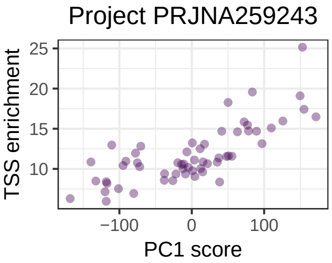


Figure 3.5 Correlation between PC1 and TSS enrichment in project PRJNA259243. Each point represents one library.

statistical power to call peaks [143]. A few metrics show such high correlation that they may be considered somewhat redundant for the purposes of standard QC; e.g., the number of peaks unsurprisingly shows very high correlation with peak territory (the amount of the genome covered by peaks). The ataqv software includes an option to output a reduced set of QC metrics by pruning out several metrics that tend to show very high correlation with other metrics. Ataqv metrics can be correlated with principal component (PC) scores in order to determine which characteristics of the libraries may be contributing most to the variance in the data across libraries. To demonstrate this, we performed a principal component analysis on the project with the most bulk ATAC-seq libraries from a single cell type and correlated the PC1 scores against ataqv metrics (Fig. 3.2h; we selected data from a single project and cell type because in a cross-project or cross-cell-type analysis PC1 would capture project or cell type). TSS enrichment showed the highest correlation with PC1 scores (Figs. 3.2h, S5), indicating that TSS enrichment may be a particularly important variable to examine during QC.

Ataqv metrics may be useful as covariates in downstream analysis, in part because they may reflect latent variables. For example, while examining a subset of ATAC-seq libraries from one study, we noticed that half of the libraries displayed a considerably different fragment length distribution than the other half. Through inspection of the sequencing read names we inferred the sequencing run and flowcell that each library was sequenced on and found that the median fragment lengths of each library covaried with the sequencing flowcell (Fig. 3.6a), suggesting that the QC metric was capturing a batch effect that otherwise may not have been apparent from the metadata (public metadata is frequently difficult to parse or missing altogether). Running a differential peak analysis with and without the QC metric median fragment length as a covariate, we found a robust shift towards more extreme p-values from the analysis when the covariate was included (Fig. 3.6b), which indicates increased statistical power after controlling for the batch effect.

To further demonstrate the utility of ataqv in identifying problematic variance, we used it to systematically explore two potential sources of bias. ATAC-seq experiments produce a stereotypical FLD, distinguished by many short (< 100 bp) fragments and a tail of longer (> 147 bp) fragments in multiples of the nucleosomal unit size. Because chromatin structure differs across classes of regulatory elements, different regulatory elements produce different local FLDs [23]. We therefore hypothesized that variables perturbing the FLD will systematically change ATAC-seq results. We therefore designed experiments to test the influence of two technical variables: Tn5:nuclei ratio and sequencing lane cluster density. As noted in [24], the ratio of Tn5 enzyme to nuclei number is a determining factor in the experiment FLD. Increasing this variable should shift the FLD toward shorter fragments. Sequencing lane cluster density also affects the length distribution of sequenced fragments, with high cluster density generally favoring shorter fragments [22, 87]. Importantly, while both of these variables affect the FLD, they do so in different ways. In the case that the Tn5:nuclei ratio changes, both the global (genome-wide) as well as local (locus-specific) FLDs should shift. When the sequencing lane cluster density changes, the true underlying global and local FLDs do not change; however, they are subsampled in different manners between the sequencing runs (high cluster density runs should sample more from the left-most part of the FLD than do low cluster density runs).

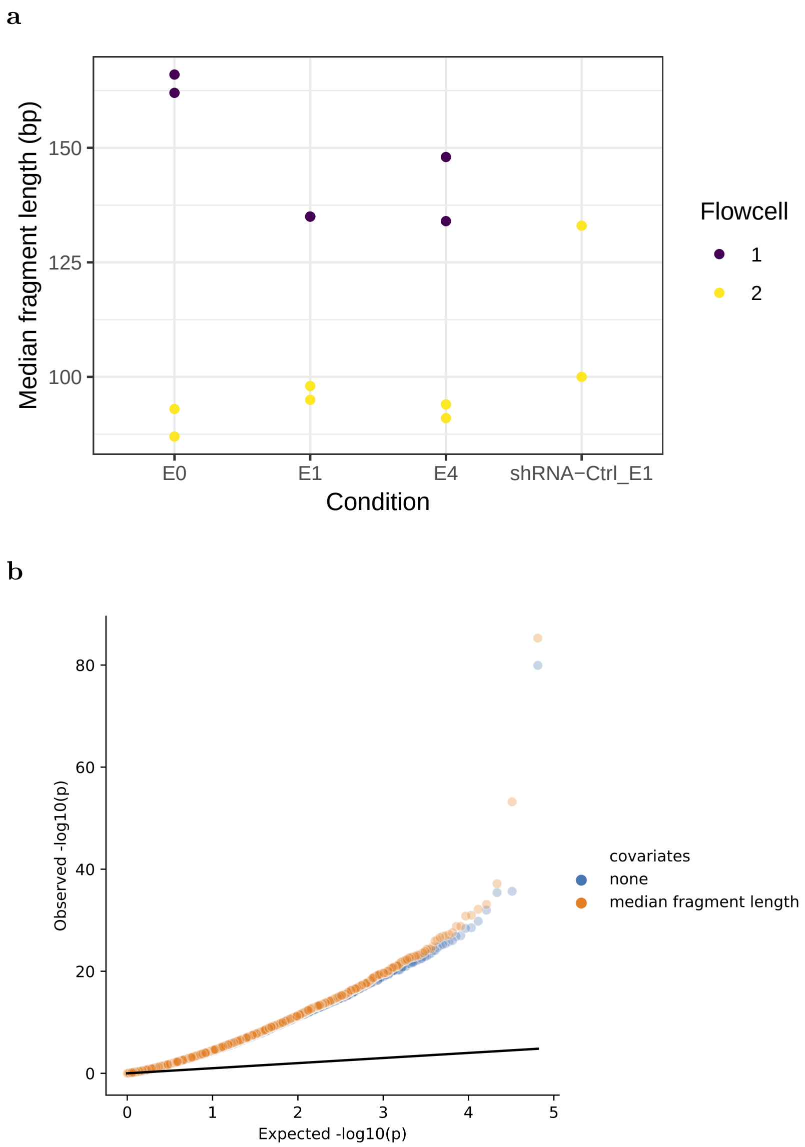


Figure 3.6 Quality control metrics may track latent variables. (A) Median fragment length vs sequencing run flowcell for a subset of libraries from project PRJNA323617/PRJNA341508. Median fragment length clearly correlates with flowcell, indicating that median fragment length may correlate with processing/sequencing batch or another related experimental variable. (B) Adding median fragment length as a covariate to a differential peak analysis between E0 and E1 results in more extreme p-values.

To quantify the influence of cluster density and Tn5:nuclei ratio, we performed two sets of ATAC-seq experiments. In one, we performed ATAC-seq on GM12878 using seven different Tn5 concentrations all using 50k nuclei as input, and sequenced each library on two separate sequencing runs, one run having 124% the cluster density of the other (411M vs 508M clusters passing filtering; Fig 3.7a; n = 3 independent nuclear isolations, producing a total of 21 libraries). Importantly, during this experiment we observed that the number of PCR cycles required for each library strongly covaried with Tn5 concentration (Fig. 3.8). Because PCR amplification can influence the fragment length distribution [74] and introduce other biases, we designed a second experiment in which we again performed ATAC-seq on GM12878 nuclei using seven different concentrations of Tn5 while holding the number of nuclei constant at 50k (n = 6 independent nuclear isolations, producing a total of 42 libraries; Fig. 3.9a, b) but additionally held the number of PCR cycles constant across libraries (Fig. 3.10). We refer to these experiments as the 'PCR-variable' and 'PCR-constant' experiments, respectively. The interactive ataqv reports for both of these experiments are available online (see section 3.5).

3.3.2 Sequencing lane cluster density biases ATAC-seq library fragment length metrics and TSS enrichment

First, we examined the effect of sequencing lane cluster density on ATAC-seq results. As expected, despite the fact that the same libraries were sequenced in both runs, the fragment length distributions from the high cluster density run were consistently shifted toward shorter fragments relative to the low cluster density run (Fig. 3.7b). The average difference in median fragment length between sequencing runs was 12 bps. Interestingly, TSS enrichment was consistently higher in the high cluster density sequencing run (average difference of 1.83; Fig. 3.7c). Other QC metrics differed consistently but to a lesser degree (see ataqv HTML report). We conclude

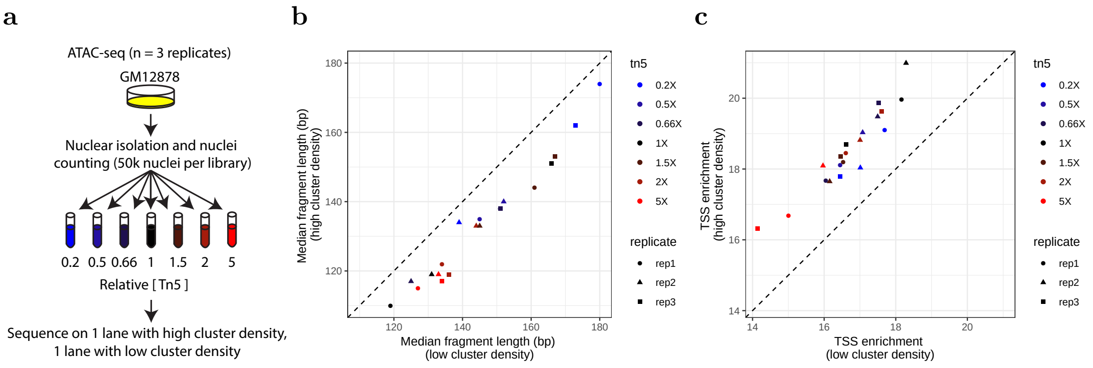


Figure 3.7 Sequencing lane cluster density systematically alters ATAC-seq results. (A) Study design for an experiment exploring the influence of sequencing flow cell cluster density. ATAC-seq was performed using 7 concentrations of Tn5; 3 replicates were used, resulting in 21 total libraries. All libraries were sequenced together in 2 sequencing runs, each on the same NextSeq 500 instrument. For one sequencing run, the flow cell loading concentration was low (about 411M clusters passed filtering), for the second the flow cell loading concentration was high (about 508M clusters passed filtering). (B) Median fragment length for each library in either sequencing run. The high cluster density run results in consistently smaller fragments, consistent with shorter fragments increasingly outcompeting longer fragments in the high-cluster-density flowcell. (C) TSS enrichment for each library using results from the high cluster density or low cluster density run. The high cluster density run results in consistently greater TSS enrichment.

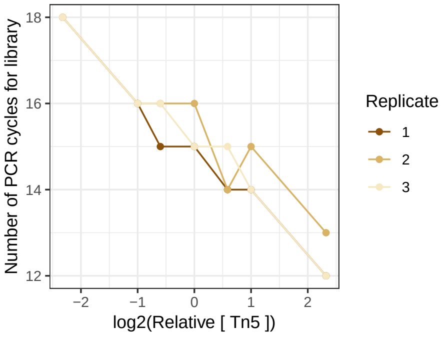


Figure 3.8 PCR cycles correlates with Tn5 concentration. Number of PCR cycles for all libraries in the PCR-variable experiment, by nuclear isolation replicate and Tn5 concentration.

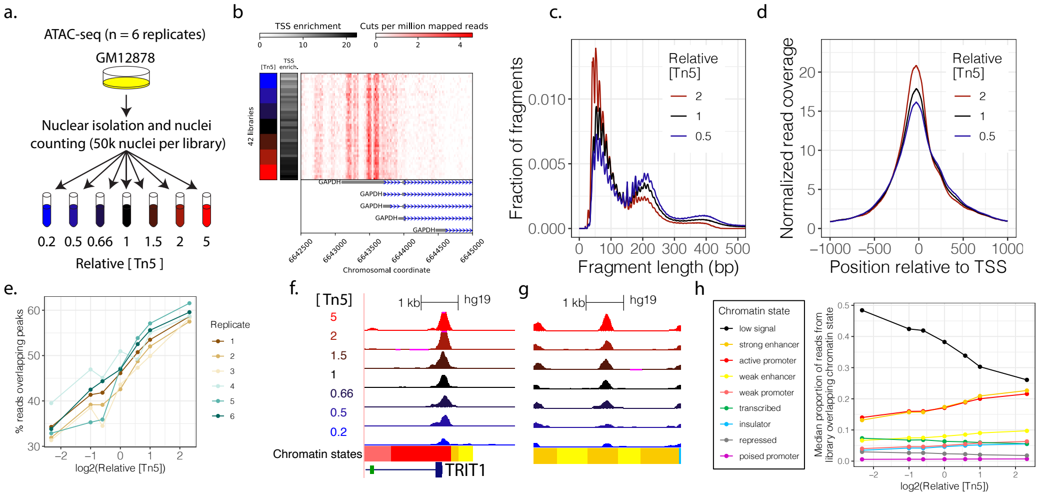


Figure 3.9 Tn5 concentration systematically alters ATAC-seq results (a) ATAC-seq was performed on GM12878 cells, using seven different concentrations of Tn5 transposase while keeping the number of nuclei constant. Six replicates were performed. (b) GAPDH locus coverage. (c) Increasing Tn5 concentration shifts the fragment length distribution towards shorter fragments. (d) Increasing Tn5 concentration increases TSS enrichment. (e) Increasing Tn5 concentration increases the percentage of high-quality, autosomal reads overlapping peaks. (f) UCSC genome browser screenshot displaying a Tn5-sensitive promoter peak (`http://genome.ucsc.edu/`) [122, 31]. (g) UCSC genome browser screenshot displaying a Tn5-sensitive enhancer peak. (h). The percentage of ATAC-seq reads falling into enhancer and active TSS chromatin states increases with increasing Tn5, while the percentage of reads falling into low signal regions decreases. Values shown represent the median values across replicates in the PCR-constant experiment.

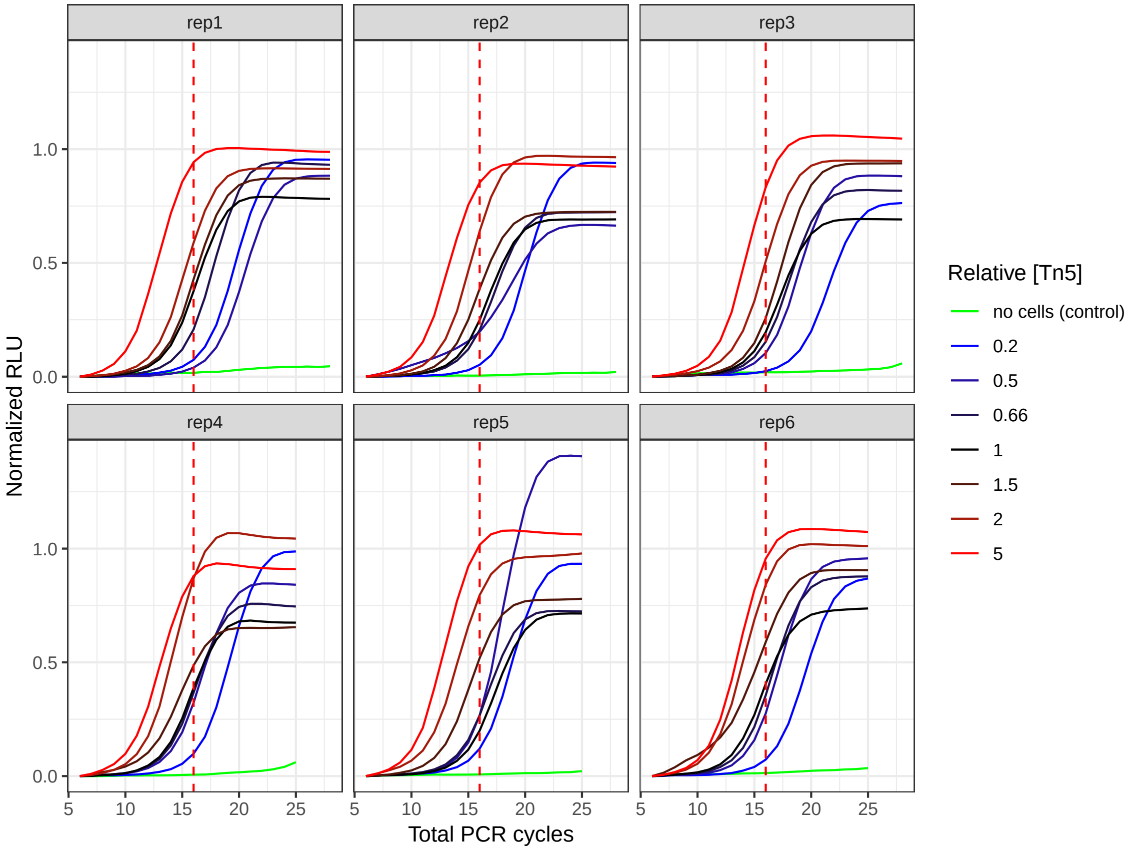


Figure 3.10 qPCR curves used for the selection of PCR amplification cycles in the PCR-constant experiment. Facets represent nuclear isolation replicates. Dashed red line represents the number of PCR cycles (16) used for all libraries.

that sequencing run cluster density has a systematic effect on ATAC-seq QC metrics, likely because different cluster densities effectively 'subsample' the actual library fragment length distribution in a biased manner and this changes the representation of different functional regions (like the TSS) across the genome.

3.3.3 ATAC-seq results are sensitive to Tn5:nuclei ratio

Next, we examined the effect of the Tn5:nuclei ratio, using the results of the PCR-constant experiment and of the high cluster density sequencing run of the PCR-variable experiment. As expected, when the number of PCR cycles was held constant, the fragment length distribution shifted toward a greater proportion of shorter fragments as Tn5 concentration increased (Fig. 3.9c, 3.11a). This correlation was attenuated when PCR cycles were allowed to vary, likely reflecting the influence of PCR cycles on FLDs (Fig. 3.12a). Furthermore, in both experiments we found that increasing Tn5 concentration negatively correlated with the percent of mitochondrial sequencing reads (Fig. 3.11b, 3.12b). We speculate that as Tn5 concentrations increase, an increasing proportion of mitochondrial DNA (which competes with nuclear DNA for the pool of Tn5) [189] is digested to the extent that it is no longer effectively sequenced. Alternatively, it may be that an increasing proportion of nuclear genomic DNA is digested sufficiently to be effectively sequenced. As mitochondrial reads are typically filtered out during standard ATAC-seq data processing, reducing mitochondrial reads increases the amount of sequence available for downstream analysis. Read duplication rate negatively correlated with Tn5 in both experiments (Fig. 3.11d, 3.12e). Overall, increasing the amount of Tn5 resulted in a considerably greater proportion (approximately four-fold higher comparing the extremes of Tn5 concentration) of reads surviving filtering in both experiments (Fig. 3.11e, 3.12f). Additionally, we found Tn5 concentration positively correlated with the enrichment of fragments around TSSs (Fig. 3.9d) in the PCR-constant experiment but not in the

PCR-variable experiment (Fig. 3.12c). The enrichment of reads in ATAC-seq peaks increased approximately 1.75-fold from the lowest Tn5 concentration to the highest (Fig. 3.9e, 3.12d). Examining the peaks called for each library, we found that the number of peaks increases with Tn5 concentration, and that this relationship is not solely due to differences in the number of reads surviving bioinformatic filtering at each Tn5 concentration (Fig. 3.13). Furthermore, we found that as Tn5 concentration increases, peak calls become more reproducible, such that the mean Jaccard index between peak calls from two replicates increases as Tn5 concentration increases (Fig. 3.14). We performed a principal component analysis and found that the first principal component correlated with Tn5 concentration (Fig. 3.15), confirming that this technical variable has a systematic effect on ATAC-seq results.

In order to determine whether there is a subset of peaks that are Tn5 sensitive or whether Tn5 sensitivity is a shared property of all peaks, we used a negative binomial generalized linear model (GLM) to model the number of reads in an ATAC-seq peak as a function of the Tn5 concentration, controlling for replicate and using the PCR-constant experiment. At a false discovery rate (FDR) of 5%, we identified 49,989 Tn5 sensitive peaks (of 70,658 total and 62,576 for which the model converged; Fig. 3.16, top panel; 3.9f,g, Fig. 3.17, 3.18, 3.19), of which the overwhelming majority (99%) displayed a positive relationship between Tn5 concentration and peak signal (49,443 of the 49,989 5% FDR peaks). Similar results were obtained in the PCR-variable experiment (29,355 peaks significant of 43,447 that converged; Fig. 3.20). This massive shift (79.9% of converged peaks in PCR-constant experiment) indicates that Tn5 sensitivity is a common quantitative trait of peaks across the genome. Adding a covariate summarizing the fragment length distribution of each library to the GLM reduced the number of Tn5 sensitive peaks detected in the PCR-constant experiment (Fig. 3.16; of the 62,576 peaks that converged in all models, 79.9% were Tn5 sensitive at 5% FDR when using no covariate and 8.3% were sensitive after

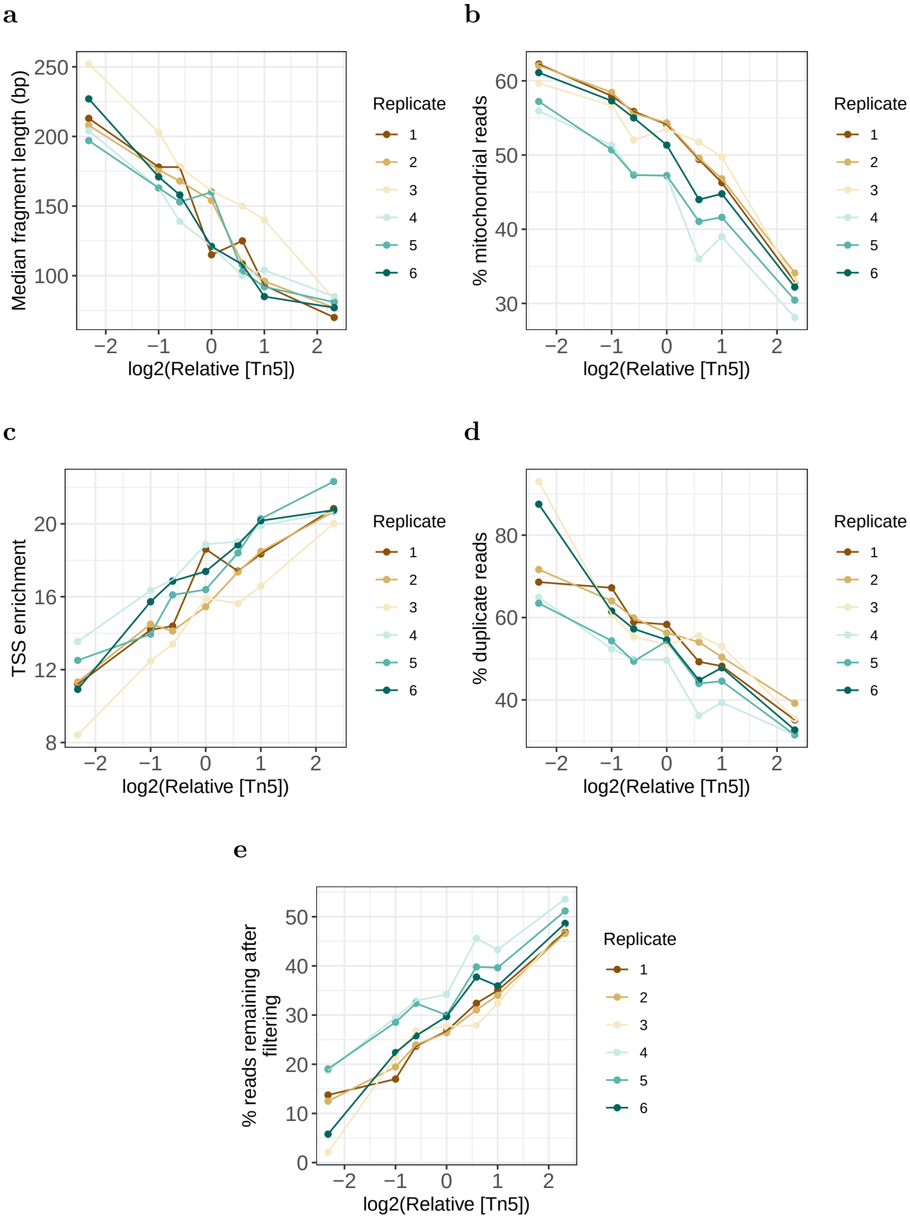


Figure 3.11 Tn5 concentration correlates with QC metrics in the PCR-constant experiment. Correlation between Tn5 concentration and (A) median fragment length, (B) percent of mitochondrial reads, (C) TSS enrichment, (D) duplication rate, and (E) proportion of reads remaining after deduplication and filtering.

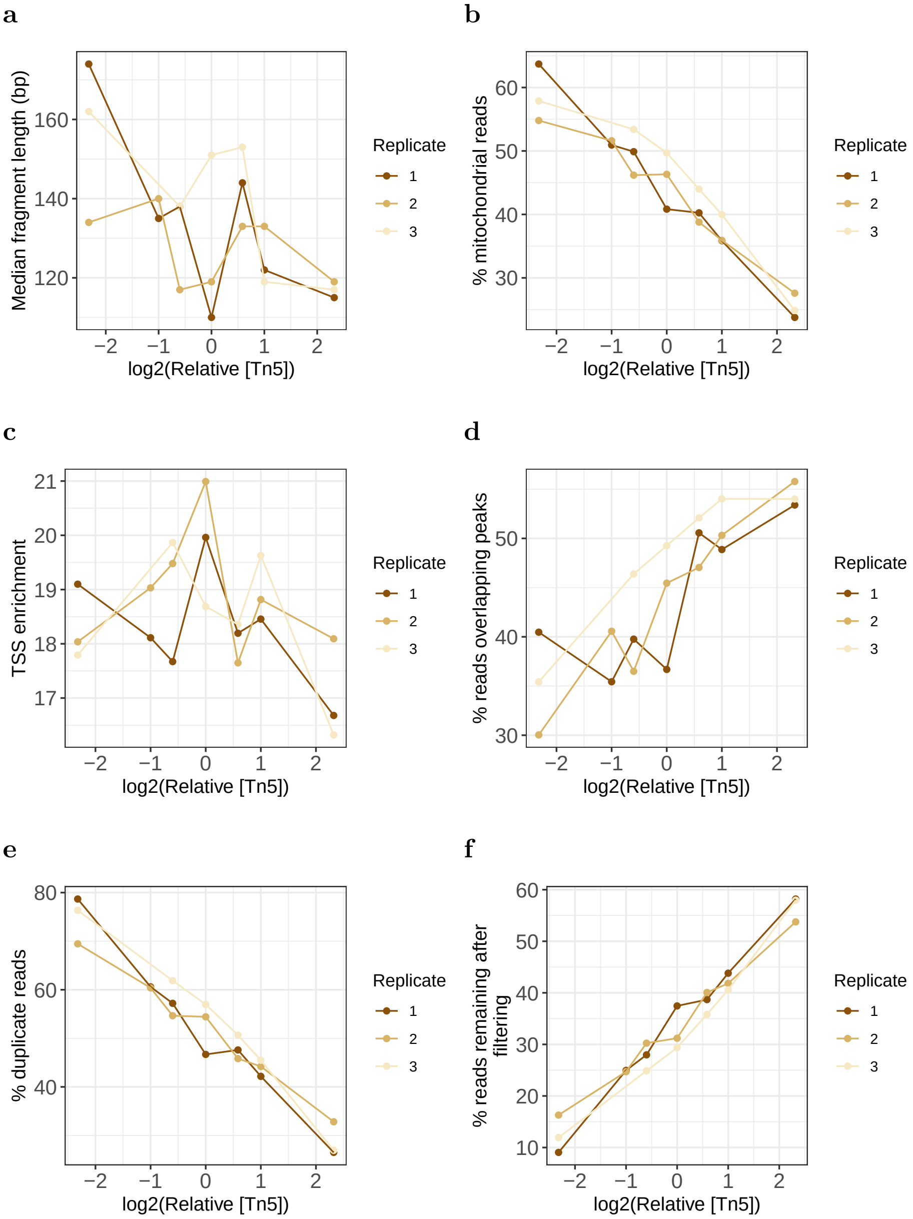


Figure 3.12 Tn5 concentration correlates with QC metrics in the PCR-variable experiment. Correlation between Tn5 concentration and (A) median fragment length, (B) percent of mitochondrial reads, (C) TSS enrichment, (D) percent of filtered reads overlapping peaks, (E) duplication rate, and (F) proportion of reads remaining after deduplication and filtering. Data shown is from the high cluster density sequencing run.

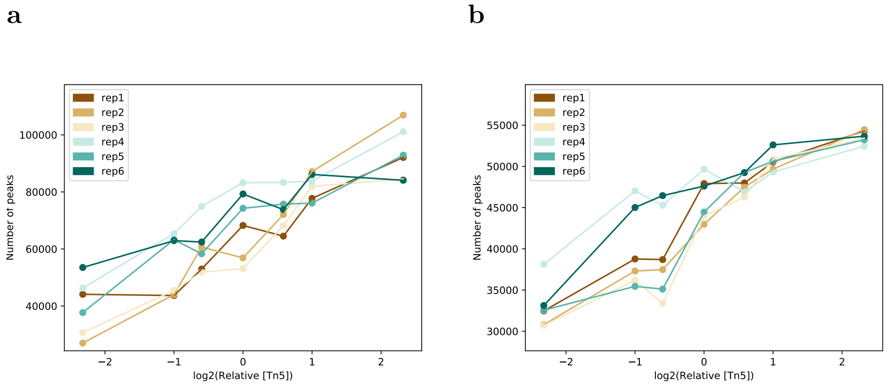


Figure 3.13 Tn5 concentration correlates with number of peaks called. Correlation between Tn5 concentration and number of peaks called (A) without and (B) with subsampling to an equal number of post-filtering reads (PCR-constant experiment).

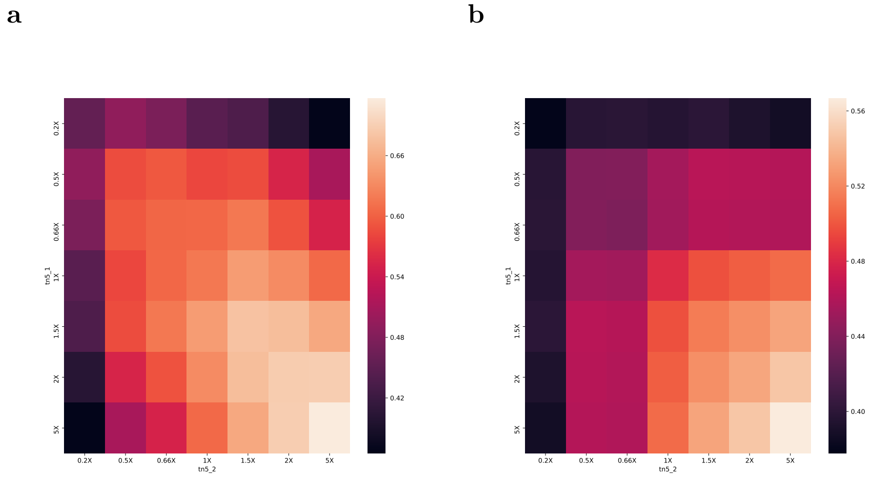


Figure 3.14 Peak call reproducibility increases with Tn5 concentration. The Jaccard index between peak calls from all pairs of libraries from the PCR-constant experiment was calculated by first taking the union of peak calls in any two libraries to generate a set of merged peaks, and then dividing the number of those merged peaks that appear in both libraries by the total number of merged peaks. A merged peak was considered to appear in a library if there was any overlap between a peak from that library and a merged peak. (A) The average Jaccard index between two replicates with given Tn5 concentrations, using all reads for peak calling. (B) The average Jaccard index between two replicates after subsampling reads to ensure that the same number of reads is used for peak calling for each library.

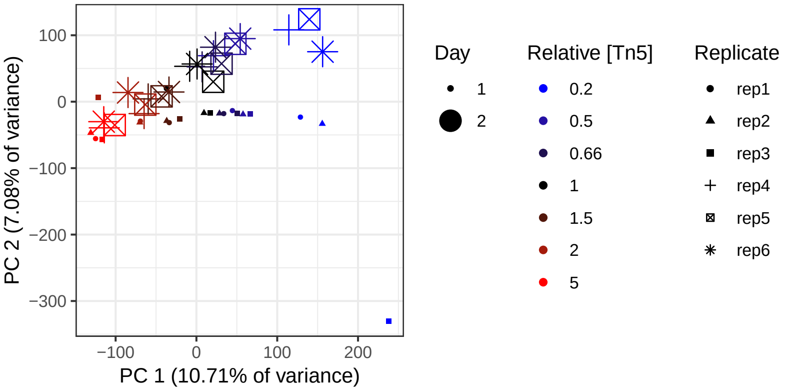


Figure 3.15 Principal component analysis of PCR-constant ATAC-seq datasets. Six replicates were used (reps 1-3 produced on day 1, reps 4-6 on day 2). The first principal component captures Tn5 concentration. The second principal component captures the day on which the experiment was performed.

adding median fragment length to the model as a covariate). Such covariates had no effect in the PCR-variable experiment (Fig. 3.20).

Our results suggest that higher Tn5 concentration increases the ATAC-seq signal-to-noise ratio (at least over the tested range of Tn5 concentrations). In order to determine if this holds for both promoter and enhancer regions, we calculated the proportion of reads that overlapped with chromHMM-derived GM12878 chromatin states (3.9f,g,h, 3.21) [67, 208]. We found that the percentage of reads falling in strong enhancer and active promoter chromatin states increases with increasing Tn5 (Bonferroni-adjusted p-values of 6.34×10^{-22} and 1.2×10^{-16}, respectively, in the PCR-constant experiment; 2.61×10^{-9} and 3.87^{-7} in the PCR-variable experiment), and that this is accompanied by a decrease in the proportion of reads falling in the low signal state (Bonferroni-adjusted $p = 1.04 \times 10^{-21}$ and $p = 1.2 \times 10^{-8}$ in the PCR-constant and PCR-variable experiments, respectively). This increase in signal-to-noise due to a technical variable is therefore observed for both TSS-proximal and TSS-distal regulatory elements.

To determine if the binding of certain transcription factors (TFs) might influence the change in ATAC-seq signal, we examined ATAC-seq reads and peaks in relation

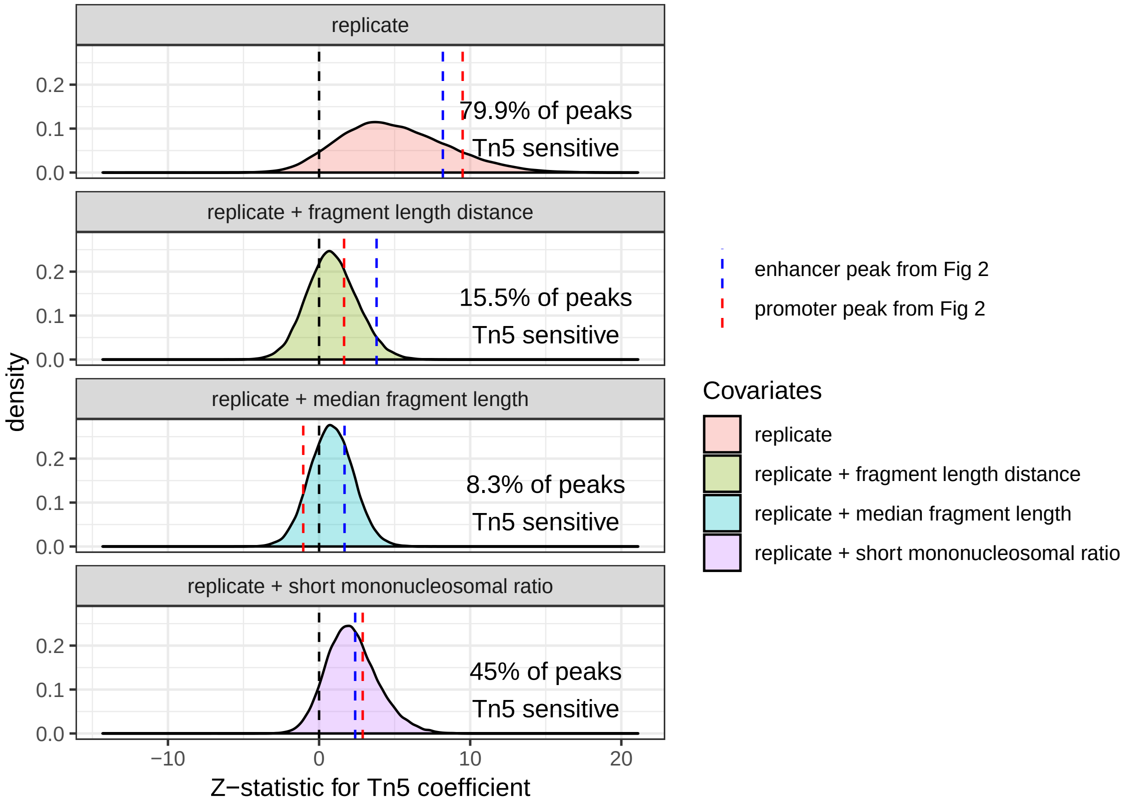


Figure 3.16 Most ATAC-seq peaks are Tn5 sensitive when PCR-cycles are held constant. Distribution of the Z-statistic for the coefficient of log2(relative Tn5 concentration) in the negative binomial GLM. The distribution is shifted in the positive direction when only the replicate is used as a covariate (top facet), indicating that many ATAC-seq peaks show increased signal (normalized to library size) as the concentration of Tn5 is increased. Adding either one of three covariates summarizing the fragment length distribution weakens the relationship between peak signal and Tn5 concentration. Percentages shown reflect only those peaks that converged in all 4 models (62,576 peaks). The example promoter and enhancer peaks from the genome browser screenshots in Fig. 2f,g are denoted by the red and blue dashed lines, respectively. They are Tn5 sensitive at 5% FDR when no FLD-summarizing variable is used as a covariate (or when short:mononucleosomal is used as a covariate), but no longer Tn5 sensitive when median fragment length or fragment length distance are used as covariates.

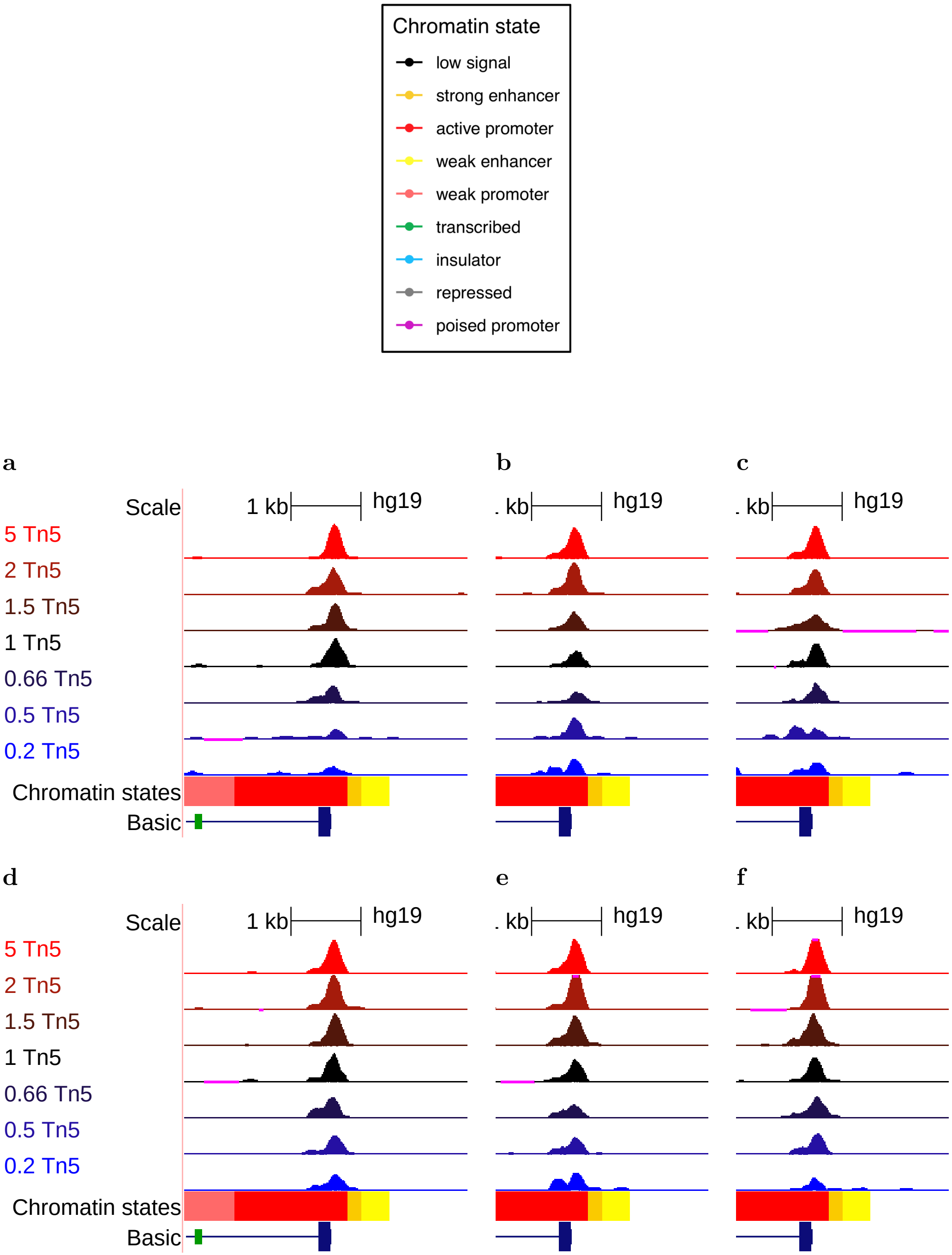


Figure 3.17 Example Tn5-sensitive promoter peak. UCSC genome browser screenshot displaying a Tn5-sensitive promoter peak in the PCR-constant experiment. The six panels represent the six replicates, displaying high reproducibility (`http://genome.ucsc.edu/`) [122, 31].

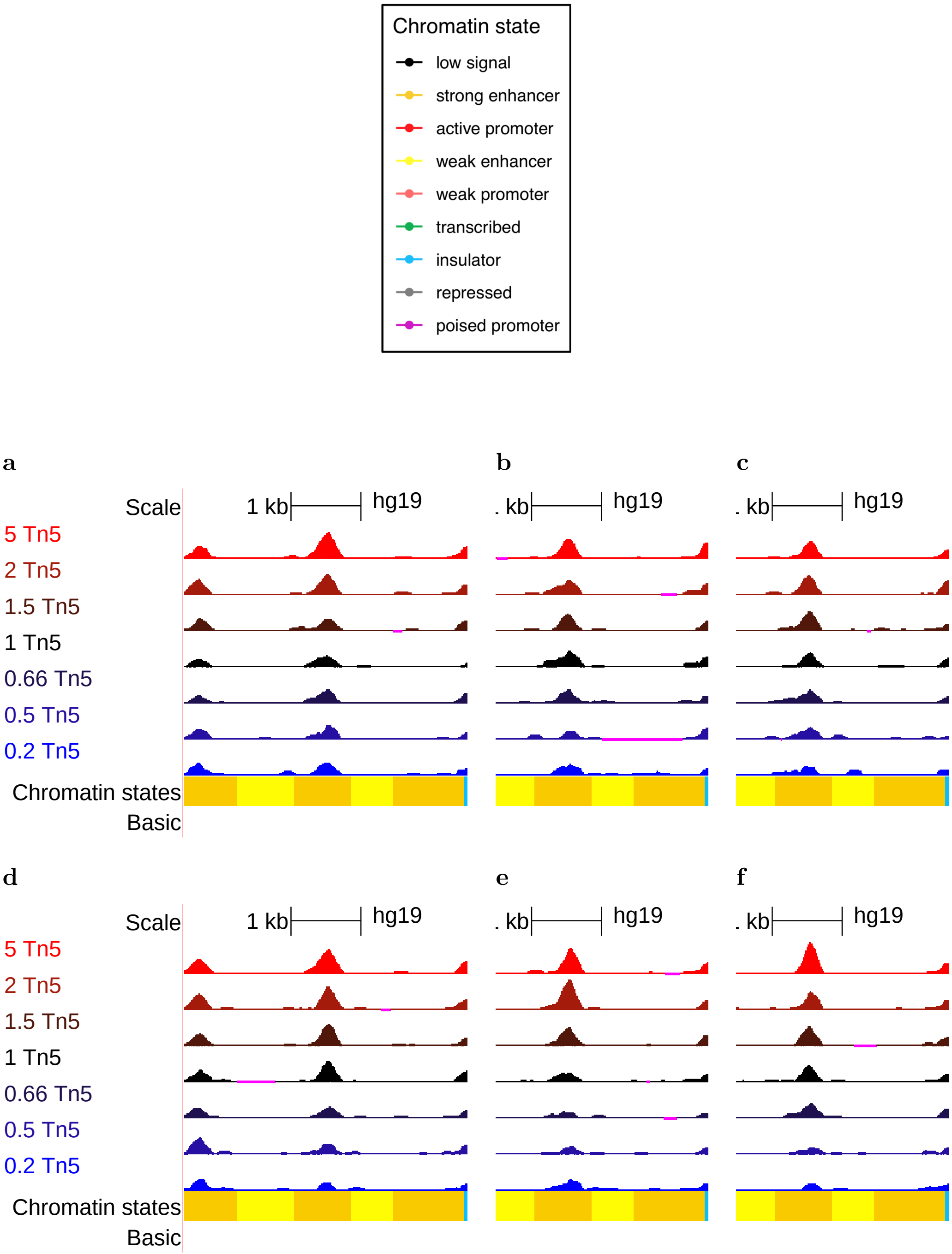


Figure 3.18 Example Tn5-sensitive enhancer peak. UCSC genome browser screenshot displaying a Tn5-sensitive enhancer peak in the PCR-constant experiment. The six panels represent the six replicates, displaying high reproducibility (`http://genome.ucsc.edu/`) [122, 31]

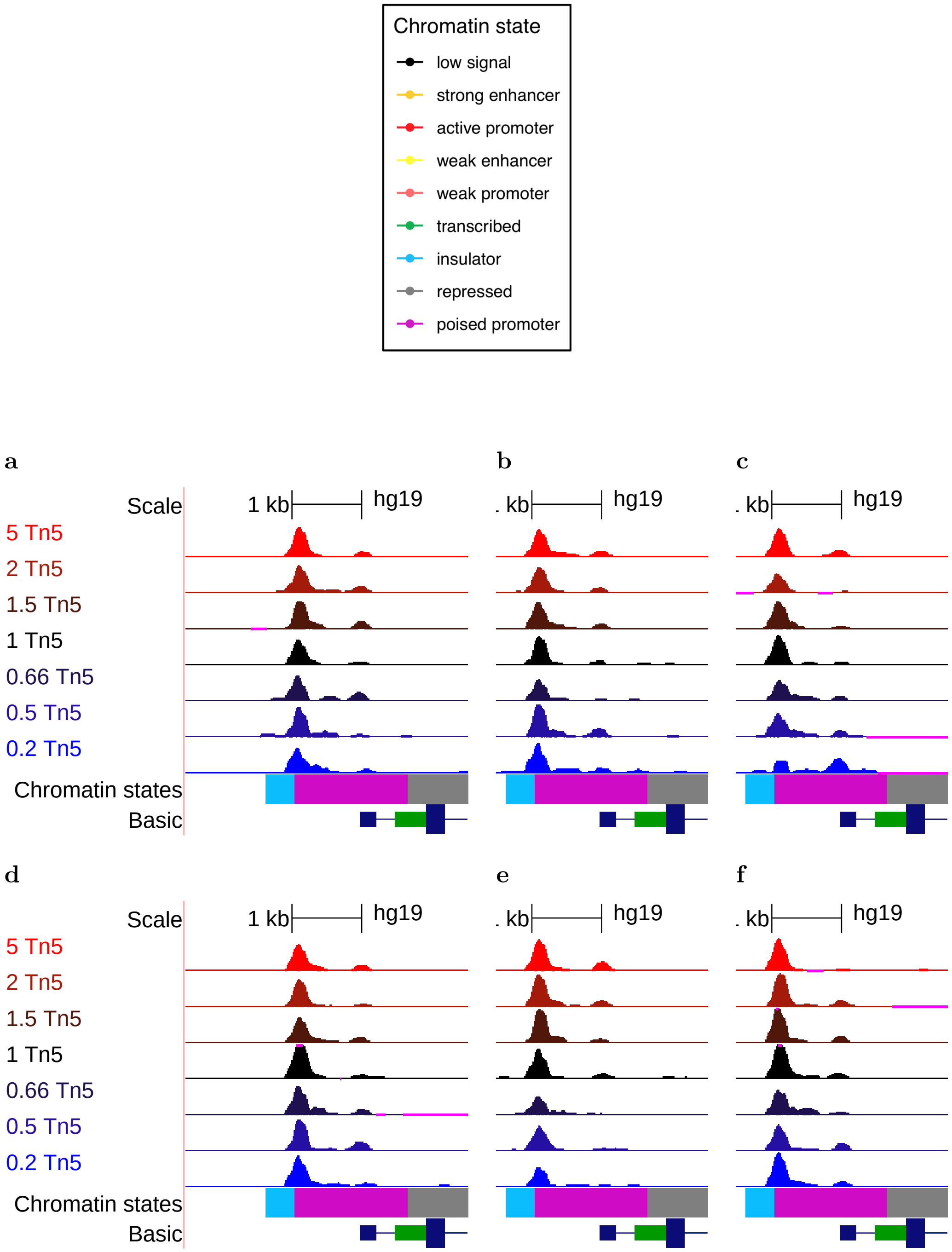


Figure 3.19 Example Tn5-insensitive peak. UCSC genome browser screenshot displaying a Tn5-insensitive peak in the PCR-constant experiment. The six panels represent the six replicates (`http://genome.ucsc.edu/`)[122, 31].

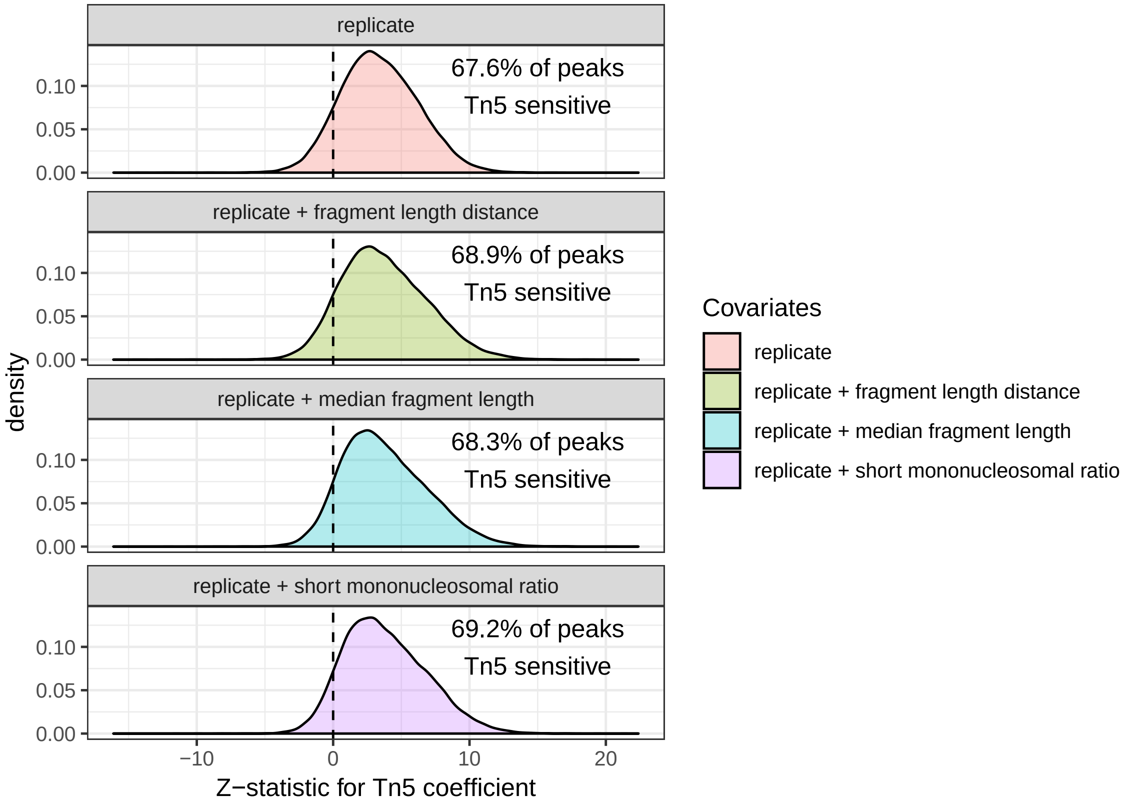


Figure 3.20 Most ATAC-seq peaks are Tn5 sensitive when PCR-cycles are allowed to vary. Distribution of the Z-statistic for the coefficient of log2(relative Tn5 concentration) in the negative binomial GLM, in the PCR-variable experiment and when including the respective covariates. The distribution is shifted in the positive direction in all cases, indicating that many ATAC-seq peaks show increased signal (normalized to library size) as the concentration of Tn5 is increased.

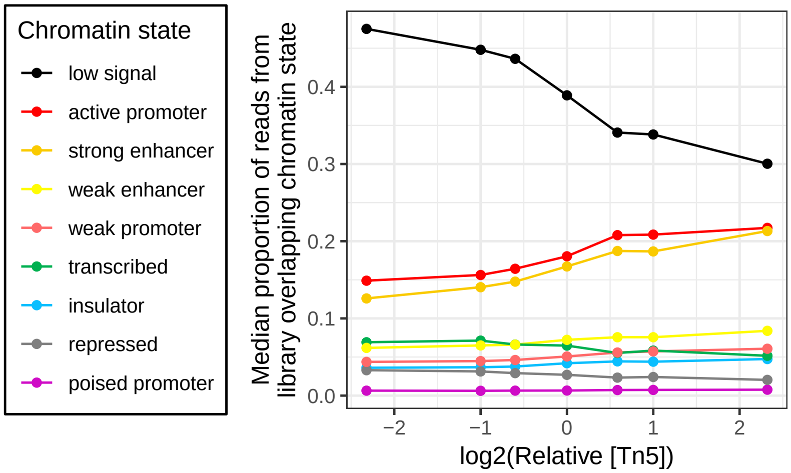


Figure 3.21 Proportion of reads overlapping with chromatin states as a function of Tn5 concentration (PCR-variable experiment). The percentage of ATAC-seq reads falling into enhancer and active TSS chromatin states increases with increasing Tn5, while the percentage of reads falling into low signal regions decreases.

to ENCODE GM12878 reproducible ChIP-seq peaks (ChIP-seq experiments on 85 TFs) [65, 267]. For all TFs, the proportion of ATAC-seq reads overlapping with ChIP-seq peaks increased as the Tn5 concentration increased (Fig. 3.22). This is consistent with our chromatin state findings (Fig. 3.9h), given that TF binding will commonly overlap with enhancers and promoters which themselves show increased signal with increasing Tn5 concentration. In order to determine if the binding of certain TFs correlates with Tn5 sensitivity, we examined the probability that a peak is Tn5 sensitive given that it is bound by a certain TF, controlling for peak size (Fig. 3.23). We performed logistic regression and discovered that binding of nearly all TFs (82 out of 85) are significantly (Bonferroni adjusted $p < 0.05$) associated with increased Tn5 sensitivity. The only exceptions are CTCF, RAD21, and REST. These factors are commonly associated with strongly phased nucleosomes [76, 97, 248, 301]; we speculate that this may render regions bound by them less sensitive to variability in Tn5 concentration. Overall, these results show that technical variation in ATAC-seq data is associated with selectively biased profiling of functional genomic regions.

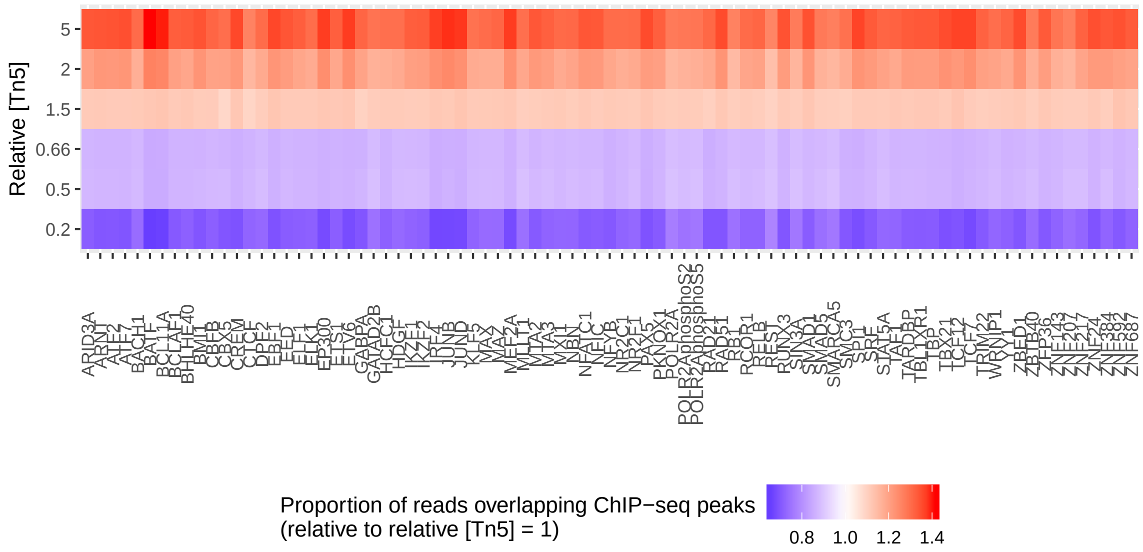


Figure 3.22 The percentage of ATAC-seq reads overlapping TF ChIP-seq peaks increases with increasing Tn5 (PCR-constant experiment). For each TF and Tn5 concentration, the mean proportion of reads overlapping ChIP-seq peaks was calculated across the six replicates before normalizing to the 1X Tn5 values.

3.4 Discussion

We conclude that ATAC-seq experiments performed for the purpose of identifying enhancers and promoters will likely achieve better signal-to-noise with increased Tn5 concentration. We note that while this relationship holds over the 25-fold range of Tn5 concentrations we have tested, it is likely that continuing to increase Tn5 concentration beyond a certain point will begin to reduce data quality as highly accessible regions are digested to an extent that they can no longer be effectively sequenced. We have not, however, reached this concentration in the data presented here. Another important caveat is that these relationships may change when nuclei numbers are limiting. The ATAC-seq protocol published in [24] states that when ”too few” cells are used, the proportion of reads derived from inaccessible regions of the genome increases. Our data was generated using a large enough number of cells that Tn5, rather than cell number, appears to be the limiting factor in library complexity. When this is the case, we find that increasing the Tn5 concentration increases the

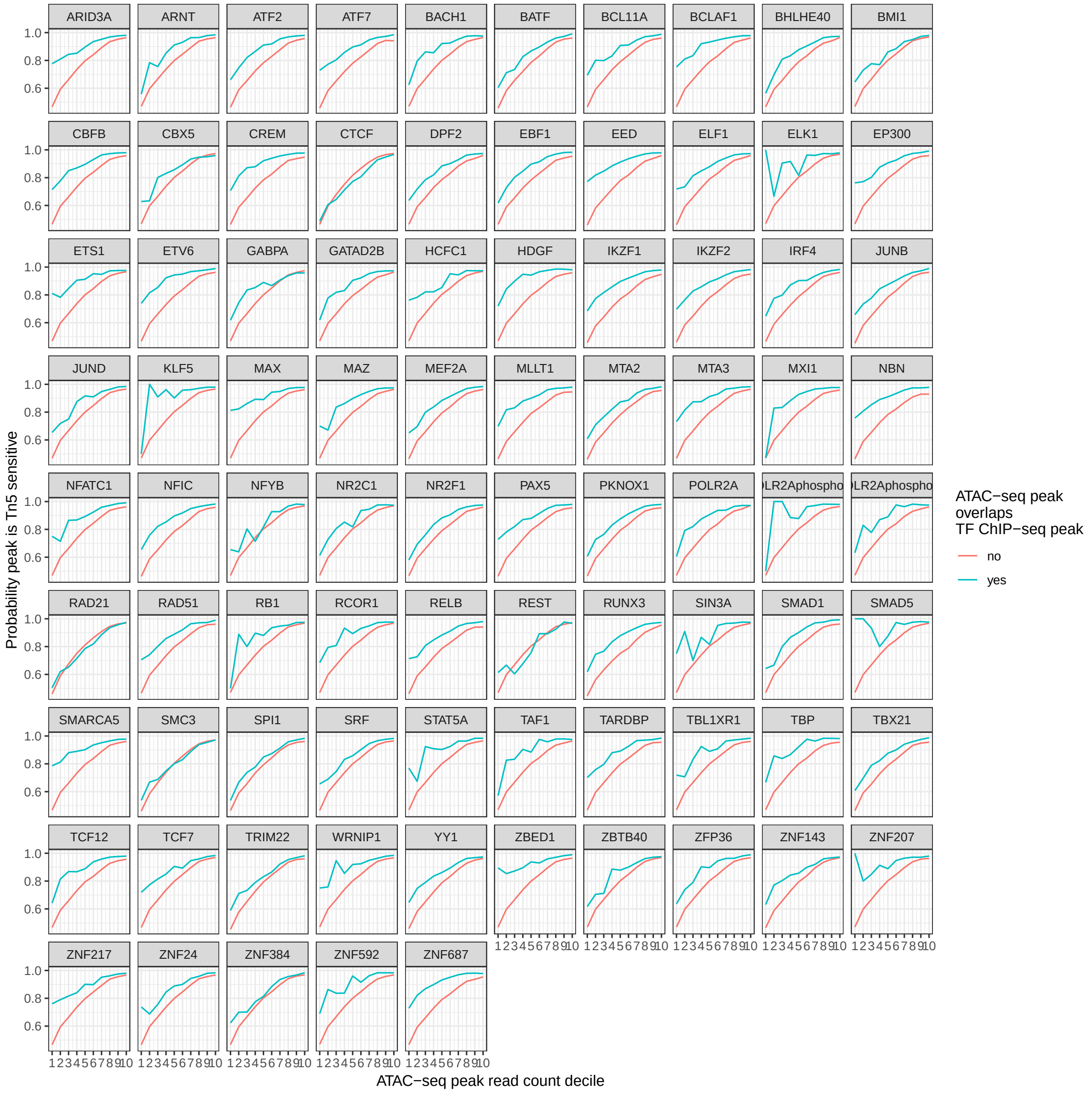


Figure 3.23 Binding of most TFs is associated with increased peak Tn5 sensitivity. Peaks are binned into deciles based on the median read count at 1X Tn5 in the PCR-constant experiment (i.e., across 6 replicates). Out of all 85 TF ChIP-seq experiments, binding of 82 TFs was significantly (Bonferroni adjusted p $\leq$ 0.05) associated with increased Tn5 sensitivity using the logistic regression approach (see section 3.5). The three exceptions are CTCF, RAD21, and REST.

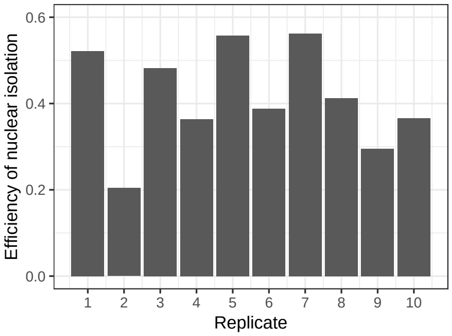


Figure 3.24 Efficiency of nuclear isolation. Nuclear isolation was performed on C2C12 cells. 250,000 cells were used as input. Efficiency is computed as (number of nuclei isolated) / (number of input cells).

proportion of reads in peaks, in enhancer and promoter chromatin states, and in most TF bound regions. These findings are generally consistent with another recent publication, which adjusted several experimental variables in cell lines and generally found that increasing Tn5 concentration yielded more peaks and greater enrichment of reads around TSS [77]; however another publication, utilizing mouse embryonic stem cells, concluded that changing Tn5 concentration had little effect on ATAC-seq results [46].

We note that, although we generated our data under a large range (25X) of Tn5:nuclei ratios, considerable differences are apparent even over lower ranges that are likely to be encountered in real-world lab settings. Differences between samples in the number of input cells and the efficiency of nuclear isolation can easily generate two-fold or greater differences in Tn5:nuclei ratio (Fig. 3.24). We expect that these differences may be especially extreme in cases of variable sample quality, or when working with tissues for which nuclear isolation is especially difficult (e.g., adipose tissue). Accordingly, nuclei counting should be a standard step in the ATAC-seq protocol to ensure consistent results.

The observed relationship between Tn5:nuclei ratio and PCR cycles is also an important finding. When the ratio is low, additional PCR cycles may be necessary in order to further enrich for short fragments in the library. This is another reason to control the Tn5:nuclei ratio, as failure to do so may lead one to differentially amplify libraries later in the protocol, which may introduce additional PCR-related biases.

Another recent publication [87] found considerable differences in the fragment length biases of different Illumina sequencing machines, and flagged this as a point of concern for those performing ATAC-seq. Our results build on and extend these published results, as we find that cluster density differences across runs on the same type of sequencing machine systematically perturb fragment length as well. Therefore, both the type of sequencing machine and the loading concentration of sequencing libraries should be taken into account when planning and analyzing ATAC-seq experiments.

When performing QC before proceeding with data analysis, a common question involves which QC metrics to focus on and which thresholds to select. We urge caution in following such hard-and-fast rules for several reasons. First, QC metrics may differ systematically according to factors like cell type. For example, embryonic stem cells are thought to have greater genome-wide chromatin accessibility than more differentiated cells [10, 155] which could result in a significantly different distribution of ATAC-seq reads and therefore differences in TSS enrichment, number of peaks, percent of reads in peaks, etc. Similarly, cell types vary in their mitochondrial DNA copy numbers [121, 273] which may lead to different levels of mitochondrial reads in different cell types, and sample heterogeneity (e.g., a homogeneous cell line vs a tissue sample composed of several different cell types) likely affects many of these metrics. Second, ideal QC metrics may depend on analysis goals. For example, if one wishes to map precise nucleosome positions adjacent to open chromatin using a method such as NucleoATAC [254], a mix of shorter and longer reads are favorable, and therefore a library with very high TSS enrichment but short median fragment length (few reads longer than 150 bps) might be considered a poor library for this purpose. These study-specific goals are therefore different, and the associated QC metrics that indicate 'good' may not be shared. Third, one's threshold for "acceptable" data will realistically vary depending on sample availability and analysis needs. If working with valuable clinical samples or very rare, hard-to-obtain cell types, the amount of material per sample or the number of samples available may be a limiting factor. In this case, one may settle for relatively lower-quality data than one would accept if one were creating abundant cell line ATAC-seq data (for which sample availability is not likely to be an issue). Lastly, we have found that the details of the calculation of a metric can make a significant difference in the resulting QC values. The calculation of TSS enrichment is a prime example. A variety of methods for calculating TSS enrichment exist among the QC packages and pipelines available. Ataqv calculates coverage around the TSS using entire ATAC-seq fragments, while other packages calculate coverage using only the cutsite or by shifting and extending individual sequencing reads such that the reads are centered on the cutsite. We have found that different methods can result in considerably different TSS enrichment values for the same library (Fig. 3.25). Unsurprisingly, the TSS list used for calculation of TSS enrichment can change results as well (Fig. 3.26) [46]. Given all of the above factors, we believe that when selecting QC thresholds researchers should look at the distribution of many QC metrics calculated uniformly across libraries, and use those distributions to determine reasonable thresholds. To demonstrate this and provide one point of reference to users, we have plotted the distributions of several QC metrics in the different cell types from the analyzed public bulk ATAC-seq data (Fig. 3.27). Similarly, if researchers wish to compare the characteristics of their ATAC-seq libraries to previously-generated libraries, a suitable reference library is probably one that is species- and cell type-matched, was processed using the same genome annotations, and that has already been shown to give quality results in the downstream analyses that the author(s) plan to utilize the newer libraries for. The ataqv packages facilitates easy implementation of all these considerations.

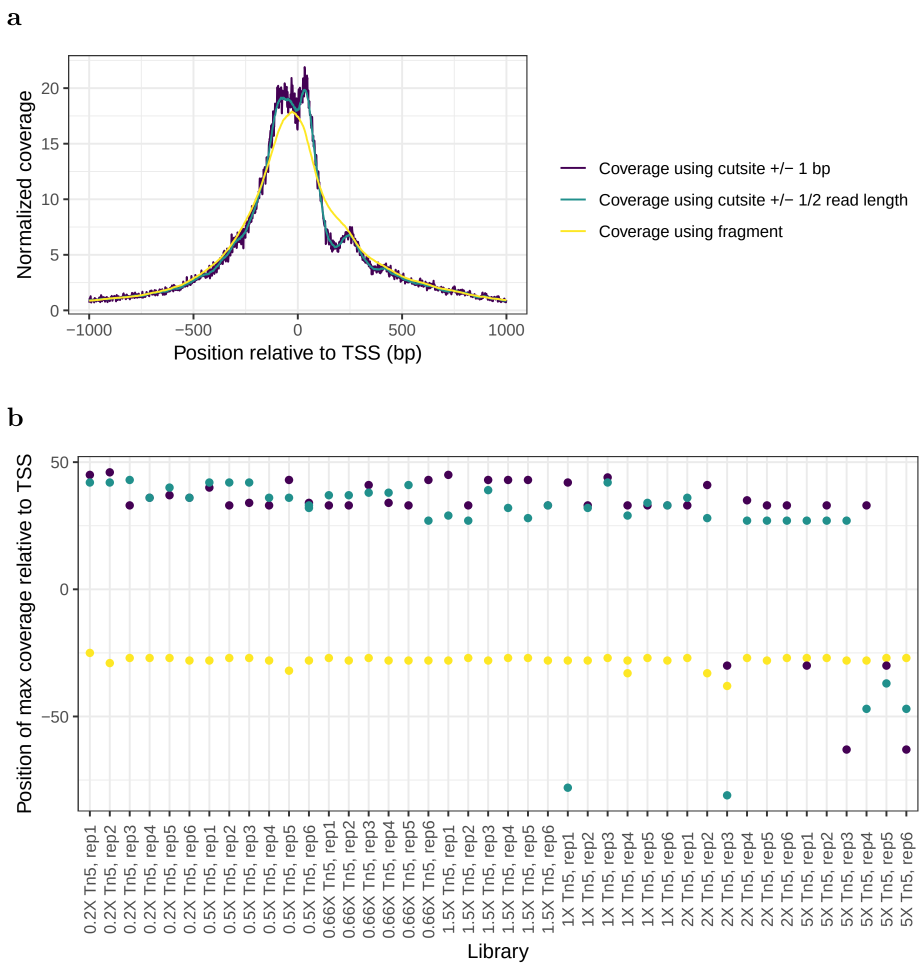


Figure 3.25 TSS enrichment curves and values are highly sensitive to calculation method. (A) Aggregate coverage around TSS for one library from the PCR-constant experiment, using different methods for calculating coverage. In one method (‘coverage using fragment’), the coverage over each bp around the TSS is calculated using the actual entire ATAC-seq fragment. In a second method (‘coverage using cutsite +/- 1bp’), coverage is calculated by determining ATAC-seq cutsites (the exact ends of each ATAC-seq fragment) and placing a fake 3-bp read over each cutsite and calculating TSS enrichment using these fake reads. In another method (‘coverage using cutsite +/- 1/2 read length’), cutsite-centered fake reads are again utilized, but the length of these fake cutsite-centered reads are set to the read length. For all methods, ‘normalized coverage’ is calculated by dividing the coverage of each bp by the average coverage over the flanking regions (-1000 - -900, and 900 - 1000). TSS coverage shows clear differences in shape and smoothness of the curve. (B) Position of max coverage (around TSS) for libraries from the PCR-constant experiment, with coverages calculated as in (A). The maximum point on the curve shows greater stability across the libraries when calculating coverage using the ATAC-seq fragment rather than cutsite-centered coverages.

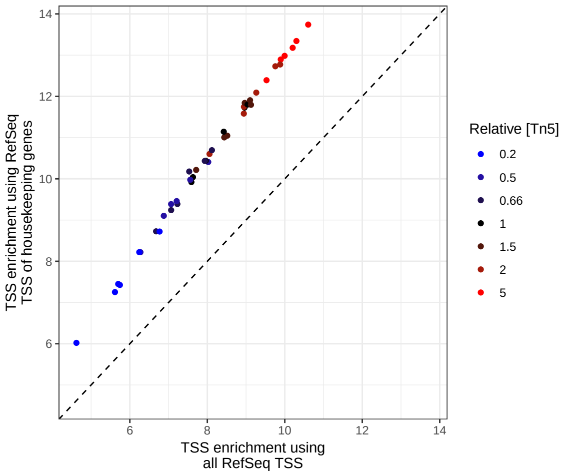


Figure 3.26 TSS enrichment is highly sensitive to TSS annotations used. TSS enrichment calculated using all RefSeq TSS or TSS of housekeeping genes only (PCR-constant experiment). Using housekeeping genes only results in systematically larger TSS enrichment values.

The systematic relationships between technical variance and change in ATAC-seq signal highlighted here demonstrate the importance of identifying and adjusting for heterogeneity in ATAC-seq data. The heterogeneity of the data may also inform one's choice of downstream methods. For example, several existing methods leverage the characteristic ATAC-seq fragment length distribution to call peaks [277], predict TF binding [159], or determine nucleosome positioning [254]. Cross-sample heterogeneity in FLDs may confound such analyses.

It has become increasingly clear that rigorous analysis of quantitative chromatin signatures will be critical for understanding complex human traits and diseases [4, 126, 139, 285]. We expect ataqv to be useful for scrutinizing confounding heterogeneity and it will therefore be an important tool in dissecting biological mechanisms.

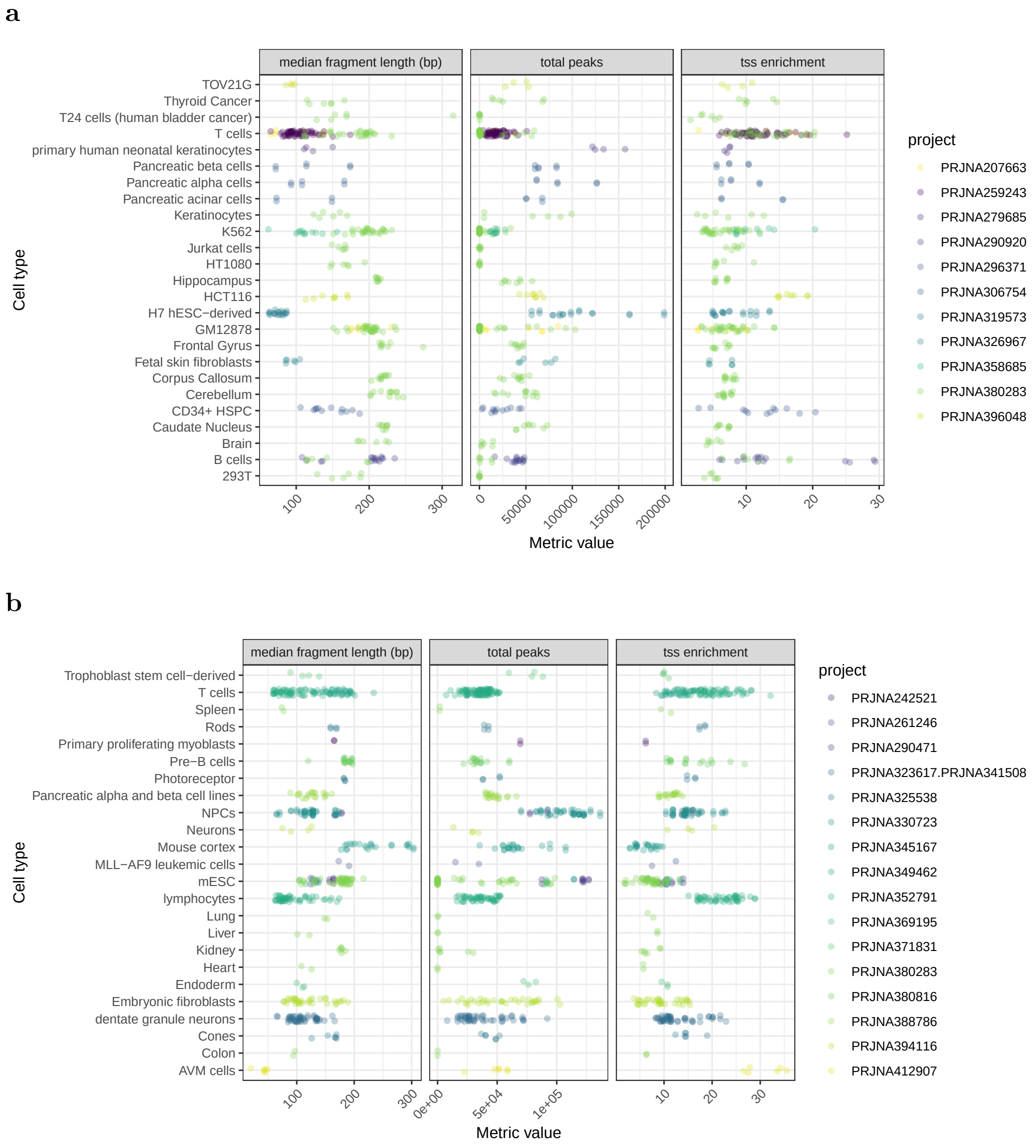


Figure 3.27 Distribution of several QC metrics across mouse and human cell types. Median fragment length, TSS enrichment, and the number of peaks in the analyzed public bulk ATAC-seq data from (A) human and (B) mouse. Each point represents one library.

3.5 Methods

3.5.1 Experimental model and subject details

We cultured GM12878 cells following the ENCODE GM12878 cell culture protocol (`https://www.encodeproject.org/documents/1bb75b62-ac29-4368-9855-68d410e1963a/`), except that we added plasmocin (Invivogen, San Diego, CA; 50 µg ml^{-1}) to the growth media to prevent mycoplasma contamination. The cell line was not authenticated. C2C12 cells were proliferated at 37 °C in growth media (DMEM + 20% FBS + 1% penicillin streptomycin) in a CO2 incubator (5% CO2). The cell line was not authenticated.

3.5.2 ATAC-seq experiments

We conducted ATAC-seq as described in [24] using a home-made Tn5 that we synthesized as described in [214]. We isolated nuclei from three independent cultures ('replicates') for the 'PCR-variable' experiment and six additional cultures for the 'PCR-constant' experiment. For each replicate we incubated 50,000 nuclei with various concentrations of enzyme ($\frac{1}{5}$ X, $\frac{1}{2}$X, $\frac{2}{3}$X, 1X, 1.5X, 2X, 5X; 1X corresponds to 2.5 µl of 1:1 Tn5-A/B mix) at 37 °C for 30 minutes in a 50 ul reaction. We column-purified the tagmented DNA using the Zymo DNA Clean & Concentrator-5 kit (Zymo Research, Irvine, CA). In the PCR-variable experiment, we PCR-amplified the entire eluate until amplification curve reached its mid-log phase ($\frac{1}{3}$ to $\frac{1}{2}$ of max signal; the number of PCR cycles required to reach this phase differed among groups, see Results section); whereas in the PCR-constant experiment, we amplified the entire eluate with a fixed number of PCR cycles (16) for all samples. We purified the products using SPRI beads prepared as in [244] and eluted in 20 ul of TE buffer with Tween-20 (10 mM Tris-HCl, 0.1 mM EDTA, 0.05% Tween-20, pH 8). Libraries were multiplexed and sequenced on an Illumina NextSeq 500 instrument.

3.5.3 GM12878 ATAC-seq data processing

All reads were trimmed to 36 bps using fastx_trimmer (from fastx-toolkit v 0.0.14). Adapters were trimmed using cta (v. 0.1.2; `https://github.com/ParkerLab/cta`). Reads were aligned to hg19 [142] using bwa mem (v. 0.7.15; flags: -M) [156]. For the ATAC-seq experiments that were used to observe the effect of Tn5 concentration, each library was sequenced on two sequencing runs; BAM files from the two sequencing runs were merged using samtools merge. Picard MarkDuplicates (v. 2.18.27; `http://broadinstitute.github.io/picard`) was used for duplicate removal (options: VALIDATION_STRINGENCY=LENIENT) and samtools (v. 1.7) [157] was used to filter for autosomal, properly-paired and mapped read pairs with mapping quality $\geq$ 30 (samtools view b h f 3 F 4 F 8 F 256 F 1024 F 2048 q 30). Peak calling was performed using MACS2 callpeak (v. 2.1.1.20160309; options: –nomodel –broad –shift -100 –extsize 200 –keep-dup all) [316]. Peaks were filtered against ENCODE blacklists (ENCODE Project Consortium, 2012) (downloaded from `http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeMapability/wgEncodeDacMapabilityConsensusExcludable.bed.gz` and `http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeMapability/wgEncodeDukeMapabilityRegionsExcludable.bed.gz`) using bedtools intersect (option -v; v. 2.27.1) [226]. Ataqv (v. 1.1.0) was run on the BAM files with duplicates marked, and the blacklists were passed as excluded regions. For the TSS file, we took the hg19.tss.refseq.housekeeping.ortho.bed.gz TSS file packaged with ataqv (representing TSS for genes with 1:1:1 human:mouse:rat orthologues where the human gene is a housekeeping gene [63]; GitHub commit f4b655) and further filtered the list to remove genes that had more than one TSS in human, mouse, or rat. One library from the PCR-variable experiment had very few reads ($\sim$ 0.5M in one sequencing run and $\sim$ 0.25M in the second) and was excluded from downstream analysis. For figures displaying ATAC-seq coverage, we normalized the signal to account for differences in library size and all signal track plots show the same range. Normalization was performed on the MACS2-created treat_pileup.bdg bedgraph files. The script used for normalization is available on GitHub (`https://github.com/porchard/normalize_bedgraph`; commit 82ab906; run with parameters '–to-number-reads 10000000'). The normalized bedgraph files were then converted to bigwig format using bedGraphToBigWig (v. 4) [123].

3.5.4 Public ATAC-seq data (mouse and human) processing

Libraries were processed in the same manner as the GM12878 ATAC-seq libraries (mapping to mm9 or hg19 as appropriate) [142, 192]. For mm9 we used the blacklist available at `http://mitra.stanford.edu/kundaje/akundaje/release/blacklists/mm9-mouse/mm9-blacklist.bed.gz` (downloaded on Jan. 25, 2013) and the mm9.tss.refseq.housekeeping.ortho.bed.gz TSS file packaged with ataqv, further filtered as described above for hg19.tss.refseq.housekeeping.ortho.bed.gz.

3.5.5 Determination of high-confidence peaks

To generate the list of peaks used in downstream analyses, we used bedtools merge to calculate the union of the FDR 1%, blacklist filtered peaks from libraries created using the 1X Tn5 concentration for each of the two experiments. We then kept, as master peaks, those intervals that overlapped with FDR 1%, blacklist-filtered peak calls from at least two of the 1X Tn5 libraries from that experiment.

3.5.6 ataqv metrics

Ataqv collects many common measurements of ATAC-seq results, as well as several new metrics that are illuminating when comparing experiments. These metrics are listed in Table 3.1. One of the metrics, fragment length distribution (FLD) distance, quantifies the similarity between each experiment's FLD and a reference FLD ('distance to reference distribution'; Fig. 3.28). This provides a quantitative indicator of over- or under-transposition of samples and may be used as a covariate in downstream analyses. The distance to reference distribution metric is similar to a signed Kolmogorov-Smirnov statistic, with the magnitude representing the maximum vertical difference between the empirical distribution functions of the reference distribution and the experiment's distribution. It is calculated as:

$$S = \begin{cases} max_x(F_e(x) - F_r(x)), & \text{if } max_x(F_e(x) - F_r(x)) > |min_x(F_e(x) - F_r(x))| \\ min_x(F_e(x) - F_r(x)), & \text{otherwise} \end{cases}$$

where S is the statistic, x represents a fragment length, and F_e and F_r represent the empirical distribution functions of the experiment's fragment length distribution and the reference fragment length distribution, respectively. The greater the magnitude of this metric, the less similar the experiment's FLD is to the reference FLD. A positive value indicates over-transposition relative to the reference FLD (a greater proportion of short fragments in the distribution relative to nucleosomal fragments), while a negative value indicates under-transposition relative to the reference. The interactive ataqv report includes plots of these FLD metrics, allowing for the quick visual identification of outliers.

Ataqv calculates TSS enrichment using fragments. Fragment coverage over the TSS +/- 1kb is computed, and the enrichment for each position is calculated by dividing this coverage by the average coverage over the outermost 200 bps in the 2-kb interval (100 bp upstream, 100 bp downstream).

3.5.7 Overlap of reads with chromatin states

Chromatin states were downloaded from `https://research.nhgri.nih.gov/manuscripts/Collins/islet_chromatin/hg19/ChromHMM/GM12878_chromHMM.bb` [208]. The bigBed file was converted to bed format using bigBedToBed (v. 1) [123]. Reads

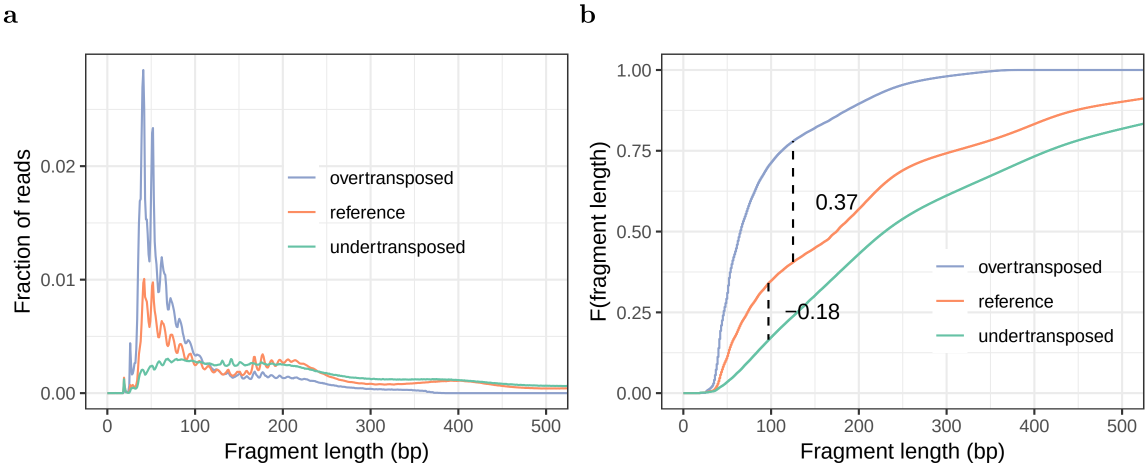


Figure 3.28 Calculation of the fragment length distribution distance. Three hypothetical distributions are displayed in panel (A); one is chosen as the reference sample. Relative to the reference sample, one sample has an abundance of shorter fragments (as would b e expected for a high Tn5 : nuclei ratio) and one has an abundance of longer fragments (as would be expected for a low Tn5:nuclei ratio). The fragment length distributions are converted to cumulative distributions (B), and the Kolmogorov Smirnov statistic is calculated (the maximum vertical distance between two cumulative distributions). The sign is then set according to whether or not the sample of interests cumulative distribution takes a value greater than (positive) or less than (negative) the reference cumulative distribution at the point of maximum vertical distance.

were filtered against ENCODE blacklist regions using bedtools intersect prior to the analysis. Each read was assigned to the chromatin state with which it showed the most overlap (according to bedtools intersect). To determine the statistical significance of the relationship between Tn5 concentration and the percentage of reads falling in each chromatin state, we ran one linear model per chromatin state, modeling proportion_of_reads_in_chromatin_state ∼ replicate + log2(relative Tn5 concentration). P-values for the Tn5 concentration coefficient were Bonferroni adjusted.

3.5.8 Overlap of reads with ChIP-seq peaks

IDR ChIP-seq peaks were downloaded from ENCODE [65, 158, 267]. Reads were filtered against ENCODE blacklist regions prior to the analysis. Bedtools intersect was used for the overlap; a single base pair was considered sufficient to call a read overlapping with a peak.

3.5.9 Estimating the efficiency of nuclear isolation

10 nuclear isolations were performed using C2C12 cells in order to characterize the variability in nuclear isolation efficiency. Cells were trypsinized and washed, and 250K cells were used for each nuclear isolation. Nuclear isolation was performed as in Supplementary Protocol 1 of [46]. For each of the 10 replicates, nuclei were counted twice using trypan blue dye in a Countess II FL instrument, and the average of the two counts used to determine the number of final nuclei.

3.5.10 Modeling Tn5-sensitive peaks

To detect Tn5-sensitive peaks, we used the glm.nb function in the MASS R package (v. 7.3-50) [288]. We used the following model:

$$Reads\_in\_peak \sim replicate + log2(relative\_Tn5\_concentration) + offset(log(size\_factor))$$

where replicate represents the nuclear isolation (of which there were 6), relative Tn5 concentration is one of (0.2, 0.5, 0.66, 1, 1.5, 2, 5), and size_factor is the total number of reads after filtering the BAM file (the 'offset' term adjusts for the variable number of reads in each library after filtering). The 'reads in peak' value was determined by passing the bed file of high-confidence peaks and each filtered BAM file to bedtools' coverageBed ('-counts' option). In the case that the model did not converge, we excluded the peak from the downstream analysis. For the PCR-constant experiment, all 42 libraries were used. For the PCR-variable experiment, the 20 libraries that passed QC were used.

3.5.11 Logistic regression to estimate TF ChIP-seq peak sensitivity

To determine if binding of each TF is associated with increased Tn5 sensitivity, we modeled

$$peak\_is\_tn5\_sensitive \sim median\_atac\_peak\_signal + overlaps\_TF\_chipseq\_peak$$

using R's glm function. The median_atac_peak_signal term controls for differences in NB GLM power as peak size increases. To calculate this term, we first gathered read counts for all Tn5 = 1X libraries in all peaks, and normalized these counts by the median count within each library to get a peak signal score for each peak in each library. We then took the median signal score across libraries for each peak. P-values for each TF were Bonferroni adjusted. Peak signal and whether or not the peak was Tn5 sensitive was derived from the PCR-constant experiment.

3.5.12 Data and code availability

All raw and processed data generated during this study have been deposited to GEO under the accession: GEO: GSE130450. We have created a GitHub repo containing the code used in this work (https://github.com/ParkerLab/ataqv-2019). Interactive QC reports for previously published ATAC-seq libraries are available at https://theparkerlab.med.umich.edu/data/porchard/ataqv-public-survey/. Interactive QC reports for our original data are available at https://theparkerlab.med.umich.edu/data/porchard/ataqv-tn5-series-pcr-controlled (PCR-constant experiment, with six replicates per Tn5 concentration) and https://theparkerlab.med.umich.edu/data/porchard/ataqv-tn5-series-not-pcr-controlled (PCR-variable experiment, and sequencing at high and low cluster density).

3.6 Acknowledgements

We thank members of the Parker laboratory for testing ataqv and recommending improvements to the design and functionality of the software. We thank Minjun Jin and Nandini Manickam for providing the nuclear isolation efficiency data. This work was supported by National Institutes of Health (N.I.H.) grant R01 DK-117960 (to Stephen C.J. Parker), the American Diabetes Association Pathway to Stop Diabetes Grant 1-14-INI-07 (to Stephen C.J. Parker), grant T32 HG00040 from the National Human Genome Research Institute of the N.I.H. (to Peter Orchard), grant T32 DK101357 from the N.I.H. (to Yoshi Kyono), and a University of Michigan Rackham Predoctoral Fellowship (to Peter Orchard).

3.7 My contributions

This project (now published in [202]) resulted from the efforts of several individuals in Stephen Parker's lab as well as Jacob Kitzman's lab. Yoshi Kyono performed GM12878 ATAC-seq experiments and I performed the computational processing and analysis of original data. John Hensley and Stephen C.J. Parker conceived the ataqv software and John Hensley programmed ataqv, to which I subsequently contributed a few minor updates and bug fixes. John Hensley and I processed public datasets. Stephen C.J. Parker and Jacob O. Kitzman conceived the GM12878 experiments and supervised the study. All authors wrote the manuscript (I contributed text to all sections).

CHAPTER IV

Human and Rat Skeletal Muscle Single-nuclei Multi-Omic Integrative Analyses Nominate Causal Cell Types and SNPs for Complex Traits

4.1 Abstract

Skeletal muscle accounts for the largest proportion of human body mass, on average, and is a key tissue in complex diseases, mobility, and quality of life. It is composed of several different cell and muscle fiber types. Here, we optimize single-nucleus ATAC-seq (snATAC-seq) to map skeletal muscle cell-specific chromatin accessibility landscapes in frozen human and rat samples, and single-nucleus RNA-seq (snRNA-seq) to map cell-specific transcriptomes in human, capturing type I and type II muscle fiber signatures, which are generally missed by existing single-cell RNA-seq data. We perform cross-modality and cross-species integrative analyses for the 30,531 nuclei, representing 11 libraries, profiled in this study, and identify seven distinct cell types ranging in abundance from 63% (type II fibers) to 0.9% (muscle satellite cells) of all nuclei. We introduce a regression-based approach to assign cell types by comparing transcription start site-distal ATAC-seq peaks to reference enhancer maps and show consistency with marker gene-based cell type assignments. We find heterogeneity in enrichment of genetic variants linked to complex phenotypes from the UK Biobank

and diabetes genome wide association studies in cell-specific ATAC-seq peaks, with the most striking enrichment patterns in muscle mesenchymal stem cells ($\sim$ 3% of nuclei). Finally, we overlay these chromatin accessibility maps on GWAS data to nominate causal cell types, SNPs, and transcription factor (TF) motifs for creatinine level and type 2 diabetes signals.

4.2 Introduction

Skeletal muscle tissue accounts for 30-40% of body mass, which is the largest tissue, on average, in adult humans and is central to basic quality of life and complex diseases [75, 108]. Like other tissues, skeletal muscle is composed of a mixture of different cell types. Most of the tissue is composed of muscle fibers, which may be categorized into different fiber types, each of which display distinct metabolic and molecular phenotypes. The proportion of muscle fibers accounted for by each fiber type varies across individuals [266]. Muscle-related diseases may differentially impact different fiber types, and fiber type proportions are associated with complex phenotypes, including aerobic and anaerobic exercise capacity and type 2 diabetes (T2D) status [275]). Muscle satellite cells are progenitors to muscle fibers, indispensable for the generation and regeneration of muscle [236]; these cells are present in skeletal muscle tissue, as are several other cell types, such as mesenchymal stem cells, that cooperate in muscle regeneration [111, 129]. Molecular associations with skeletal muscle tissue/muscle fiber characteristics and muscle-related complex diseases could be mediated in part by these stem cell-like populations; for example a genetic variant that alters the developmental of a satellite cell could carry important implications for later muscle function, just as some T2D-associated variants are proposed to impact pancreatic/beta cell development rather than the function of mature beta cells [178, 282] and facial morphology associated variants may act through progenitor cell populations [310]. Immune cells infiltrate muscle tissue and communicate with muscle cells as well, playing a particularly important role following injury [215]. Profiling the transcriptomic and epigenomic landscapes of these cell types and muscle fiber types may therefore contribute to our understanding of the biology of muscle development and muscle-related complex traits.

Bulk profiling of skeletal muscle tissue ignores this heterogeneity and is dominated by the most common cell types (muscle fibers), but single-cell/-nucleus methods overcome this and allow profiling of the constituent cell types. In the case of skeletal muscle, the distinction between single-nucleus and single-cell profiling is particularly important as (1) skeletal muscle fibers have an elongated shape that may make them difficult to capture in single-cell suspensions, and (2) muscle fibers are multinucleated, meaning that a single-cell measurement will capture the output of many nuclei. Previous single-cell RNA-seq studies of human [185, 246, 308], mouse [52, 54, 83, 200, 212, 274], and pig [222] skeletal muscle tissue either capture no muscle fiber nuclei or capture them in unrepresentative proportions. Bulk analysis of pooled, dissected muscle fibers have generated fiber-type specific transcriptional profiles [6, 33, 34, 232] and analysis of specific isolated muscle resident cell populations [39, 78, 163] have generated insights into targeted cell subpopulations but these studies are necessarily biased towards specific cell types. To date no single nucleus ATAC-seq (snATAC-seq) studies of whole human or rat skeletal muscle tissue samples has been performed.

Here, we employ single-nucleus RNA-sequencing (snRNA-seq) and ATAC-seq (sn ATAC-seq) on the 10X Genomics platform to profile gene expression and chromatin accessibility of frozen skeletal muscle cell populations in human and rat. First we examine the influence of fluorescence activated nucleus sorting (FANS) and nucleus loading concentration on the performance of the platform. Next, we perform joint clustering of the snRNA-seq and snATAC-seq libraries to determine the cell types detected in skeletal muscle tissue samples and map their respective transcriptomes and chromatin landscapes. We then integrate the resulting genomic maps with UK Biobank and T2D-related GWAS results to explore the relationship between these cell types and a broad range of human phenotypes and diseases and nominate causal SNPs at several genomic loci.

4.3 Results

4.3.1 FANS negatively impacts 10X snATAC-seq results

Before being loaded onto the 10X platform, nuclei must be isolated from the samples of interest. This process involves cell lysis, which produces viable nuclei as well as substantial cellular debris and dead nuclei, some of which inevitably remains in the final nuclei suspension. By staining the DNA in live nuclei and using FANS to selectively filter the suspension for stained entities, one should be able to remove dead nuclei and cellular debris in the suspension, improving the purity and quality of the suspension loaded onto the 10X platform. However, the FANS process could stress the live nuclei or otherwise alter the snRNA-seq and snATAC-seq results. Comparing quality control metrics and (in the case of snRNA-seq) aggregate gene expression or (in the case of snATAC-seq) aggregate ATAC-seq peaks/signal between snRNA-seq and snATAC-seq libraries generated from nuclei that either did or did not undergo FANS would allow one to detect substantial changes that FANS may introduce. Also, because the aggregate of reads from a snRNA-seq or snATAC-seq library should resemble the profile of the same assay performed in a bulk fashion on the same biological sample, one can generate bulk and single-nucleus libraries from a single sample and compare quality control metrics and gene expression/ATAC-seq signal between them. Therefore, to determine the effect of FANS on 10X snRNA-seq and snATAC-seq results, we performed three nuclear isolations from a single human muscle sample, mixed the resulting nuclei together, and performed FANS on one half of the suspension (Fig. 4.1A). The FANS and non-FANS suspensions were then each used to produce two replicate snATAC-seq and two replicate snRNA-seq libraries, resulting in eight total libraries (four snATAC and four snRNA). We also generated two independent bulk ATAC-seq libraries from the same biological sample, allowing us to compare snATAC-seq profiles, with and without FANS, to a comparable bulk ATAC-seq profile.

First we examined the four snATAC-seq libraries, comparing the aggregate signal for each library to bulk ATAC-seq libraries from the same biological sample. We called peaks for the four libraries and ran the ataqv quality control software package [202] on the aggregated data to examine the overall transcription start site (TSS) enrichment and fragment length distributions. The fragment length distributions for each library resembled the expected stereotypical ATAC-seq fragment length distribution, showing an abundance of short fragments as well as mononucleosomal fragments (Fig. 4.1B) [23]; however, the TSS enrichment was lower in the FANS libraries (Fig. 4.1C), indicating the FANS libraries had a lower signal to noise ratio. This difference in signal-to-noise ratio is demonstrated when visualizing the ATAC-seq signal at genomic regions active in muscle, such as the *ANK1* locus (Fig. 4.1D) [259]. We additionally overlapped TSS-distal ATAC-seq peaks from each of the libraries with existing chromatin states from diverse tissues and cell types [243] and found that the peaks from the non-FANS libraries showed considerable overlap with skeletal muscle enhancers, while the peaks from the FANS libraries showed poor overlap (Fig. 4.2). ATAC-seq signal across FANS libraries showed poor correlation with the two bulk ATAC-seq libraries from the same sample (Fig. 4.3). We therefore concluded that FANS has a clear negative impact on 10X snATAC-seq results.

Next we examined the four snRNA-seq libraries. All four libraries showed high correlation, indicating that FANS does not substantially alter snRNA-seq results, at least at the pseudo-bulk gene expression level (Fig. 4.1E). In order to determine if

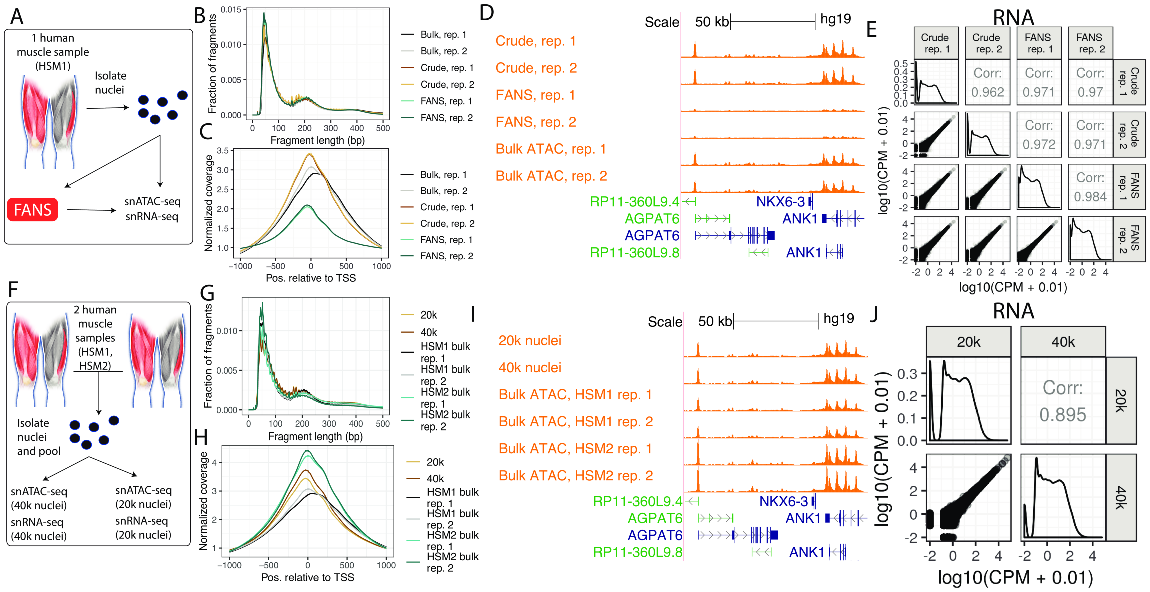


Figure 4.1 Effect of FANS and 10X well loading concentration on aggregate snATAC-seq and snRNA-seq results. (A) Study design to determine the effect of FANS on snRNA-seq and snATAC-seq results. Muscle cartoon adapted from [259]. HSM1 refers to one specific skeletal muscle sample ('human skeletal muscle 1'). Bulk ATAC-seq was performed on HSM1 as well (two replicates, each separate nuclei isolations). (B) Fragment length distribution and (C) TSS enrichment for two snATAC-seq libraries that did not undergo FANS and two that did, as well as two bulk ATAC-seq replicates from the same sample ('Bulk'). (D) ATAC-seq signal at the *ANK1* locus for FANS or non-FANS input snATAC-seq libraries, and the two bulk ATAC-seq libraries. All tracks are normalized to 1M reads. (E) Correlation between FANS and non-FANS snRNA-seq libraries; each point represents one gene. (F) Study design to determine the effect of loading 20k vs 40k nuclei into the 10X platform, utilizing HSM1 as well as a second sample, HSM2 ('human skeletal muscle 2'). Bulk ATAC-seq was performed on HSM1 (same libraries as in (a)) and on HSM2 (two replicates, each separate nuclei isolations). (G) Fragment length distribution and (H) TSS enrichment for snATAC-seq libraries after loading 20k vs 40k nuclei, as well as for the four bulk ATAC-seq libraries (two each from the two muscle samples, 'HSM1 bulk' and 'HSM2 bulk'). (I) ATAC-seq signal at the *ANK1* locus for the 20k and 40k libraries and the four bulk ATAC-seq libraries. All tracks are normalized to 1M reads. (J) Correlation between snRNA-seq libraries resulting from loading 20k vs 40k nuclei.

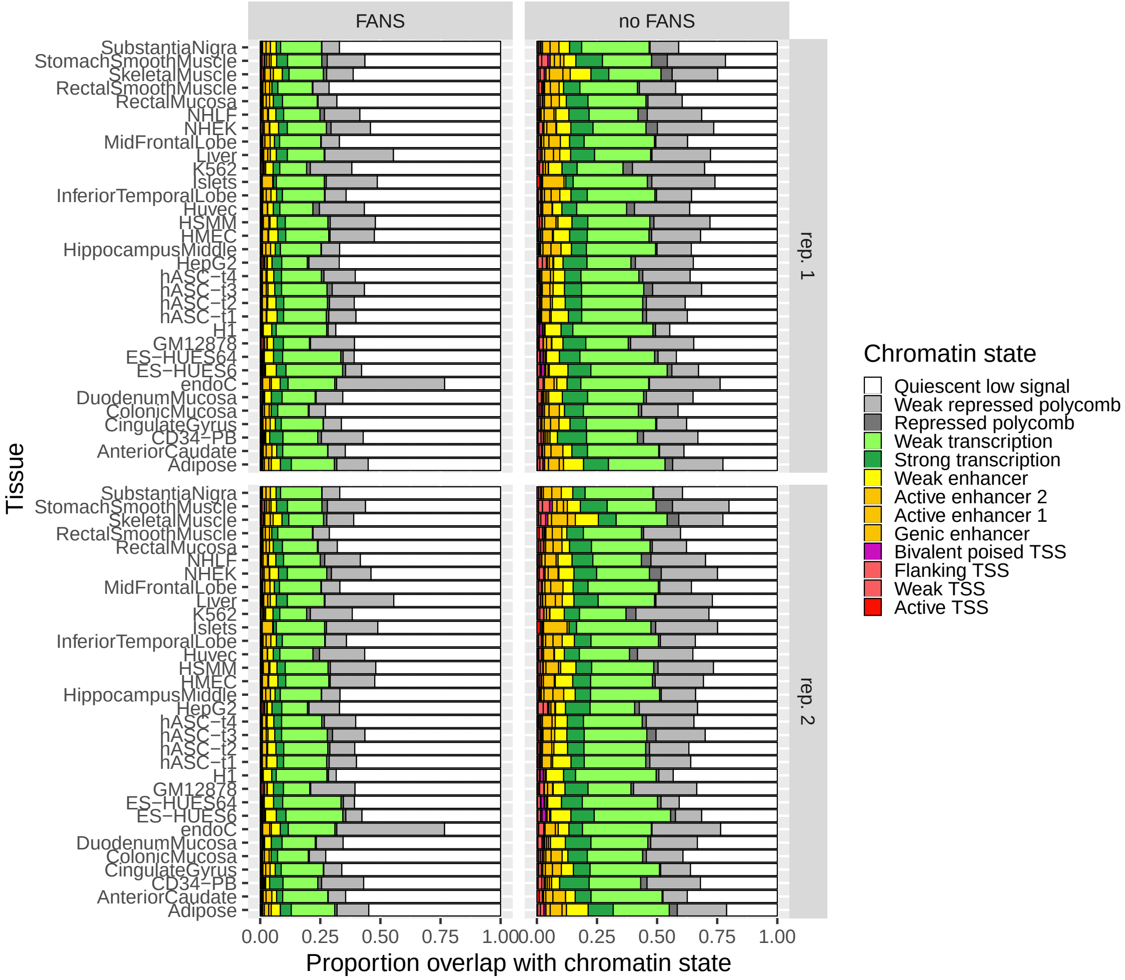


Figure 4.2 Chromatin state overlap for TSS-distal (> 5kb from TSS) ATAC-seq peaks from the FANS and non-FANS snATAC-seq libraries.

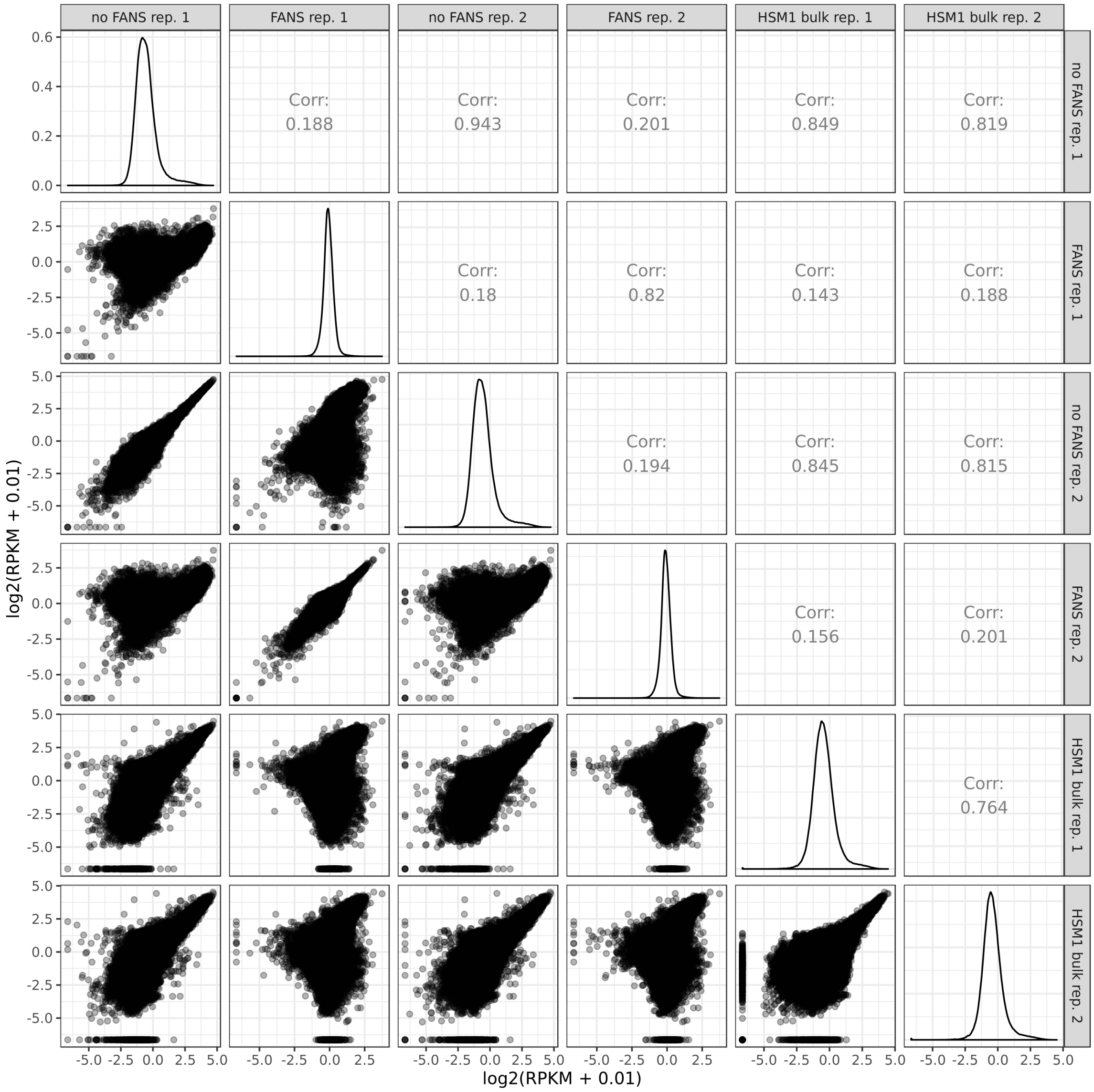


Figure 4.3 Correlation between FANS snATAC-seq, non-FANS snATAC-seq, and standard bulk ATAC-seq libraries. Each point represents one peak.

FANS altered the yield of quality nuclei, we used read counts and mitochondrial contamination to select quality nuclei from each library, additionally removing doublets using doubletfinder [181]. We found that FANS substantially increased the number of quality nuclei obtained (2,004 and 2,078 for non-FANS libraries; 7,715 and 7,118 for FANS libraries). We therefore concluded that FANS has little effect on pseudo-bulk gene expression measurements, but may alter nucleus yield.

4.3.2 snATAC-seq and snRNA-seq results are robust to nucleus loading concentrations

The concentration at which nuclei are loaded onto the 10X platform is an important parameter affecting data quality and the number of nuclei available for downstream analysis. Increasing the loading concentration increases the maximum number of nuclei from which data can be obtained; however, it also increases the probability that multiple nuclei end up with the same gel bead, thereby increasing the doublet rate. Balancing these outcomes is important to maximize the amount of quality data and number of nuclei available for downstream analysis. To evaluate the effect of increasing the number of nuclei loaded onto the platform, we performed a separate experiment in which we isolated nuclei from two muscle samples, mixed them together, and then loaded either 20k or 40k nuclei (as quantified by a Countess II FL Automated Cell Counter) into a 10X well for snRNA-seq and for snATAC-seq (Fig. 4.1F). We also generated two independent bulk ATAC-seq libraries from the biological sample for which bulk ATAC-seq profiles were not already available, allowing us to compare snATAC-seq profiles to comparable bulk ATAC-seq profiles.

The snATAC-seq libraries displayed the expected fragment length distributions and comparable TSS enrichments (Fig. 4.1G, H). We examined the aggregate signal of the snATAC-seq libraries next to bulk ATAC-seq libraries from the same samples and confirmed that both libraries showed strong signal, comparable to that of bulk data

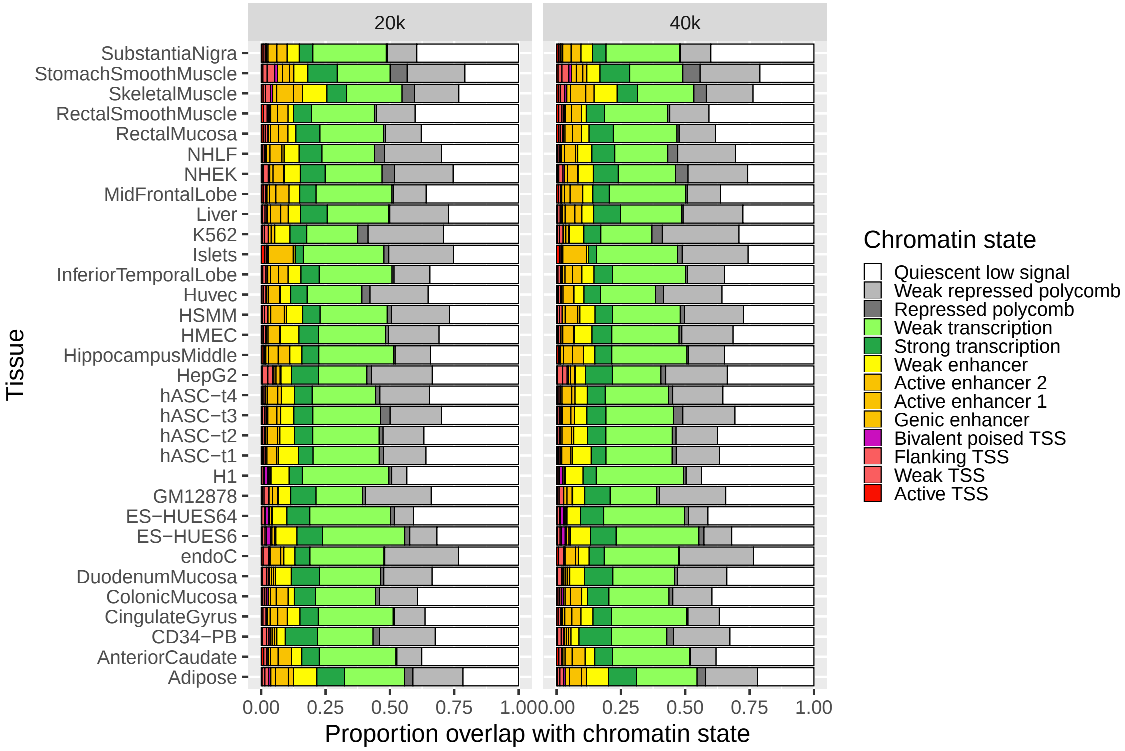


Figure 4.4 Chromatin state overlap for TSS-distal (>5 kb from TSS) ATAC-seq peaks from the 20k and 40k nucleus FANS snATAC-seq libraries.

(Fig. 4.1I). Overlap between TSS-distal ATAC-seq peaks called on both libraries and chromatin states were likewise similar, showing relatively high overlap with skeletal muscle enhancers (Fig. 4.4), and the ATAC-seq signal in the libraries correlated with bulk ATAC-seq signal to an extent comparable to the correlation between two bulk ATAC-seq libraries (Fig. 4.5). After selecting quality nuclei (Fig. 4.6), we found that the higher loading concentration yielded 2,035 nuclei while the lower concentration yielded 855 nuclei (after doublet removal).

Correlation between the snRNA-seq libraries was high, indicating that the loading concentration could be changed substantially without compromising data quality (Fig. 4.1J). We again found the higher loading concentration yielded more quality nuclei than the lower concentration (3,839 vs 2,118) after doublet removal.

10X guidelines recommend loading 15k nuclei into a well; however, our results indicate that exceeding this loading concentration can still yield quality snATAC-seq

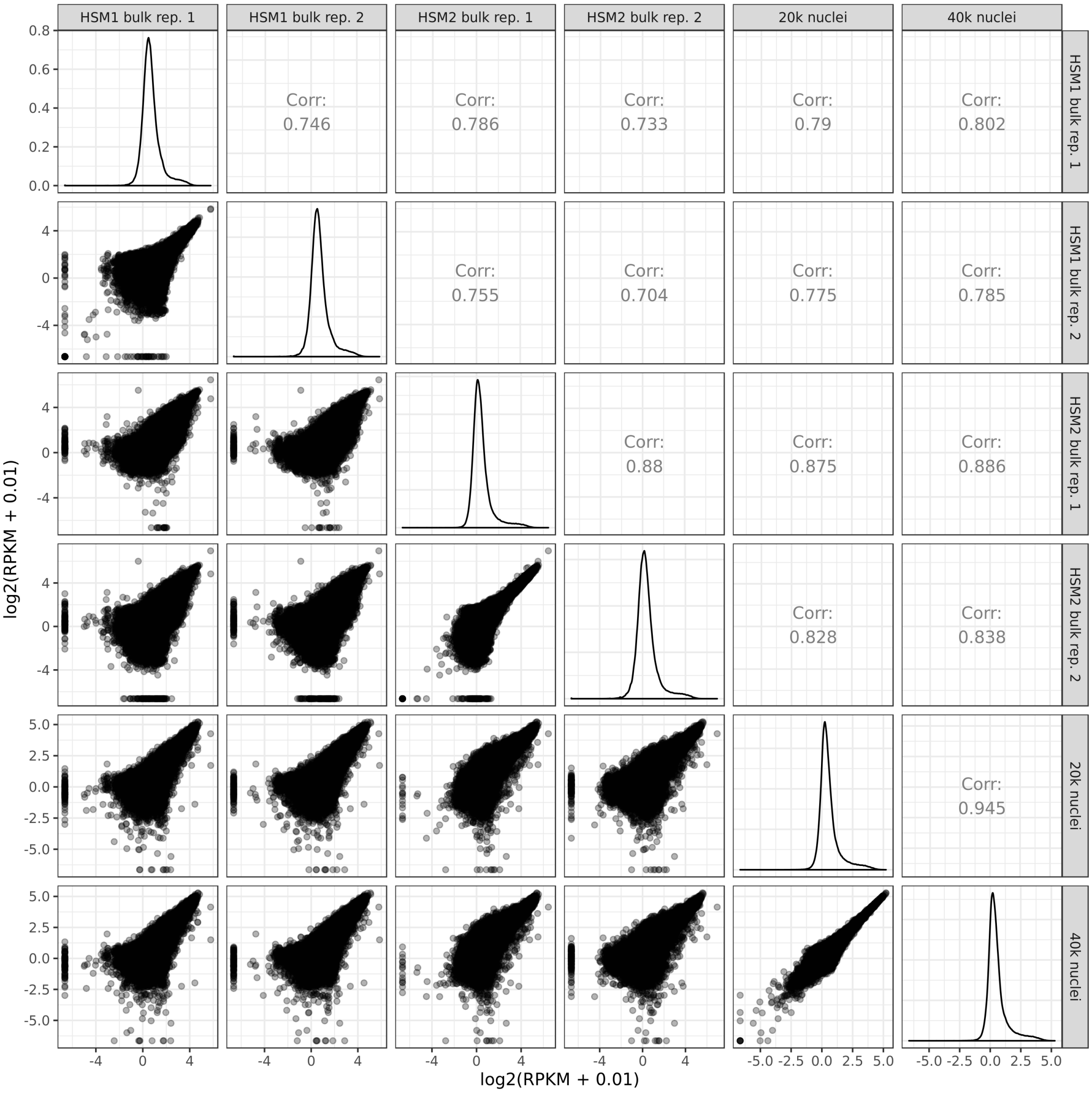


Figure 4.5 Correlation between 20k and 40k nucleus snATAC-seq libraries and standard bulk ATAC-seq libraries. Each point represents one peak.

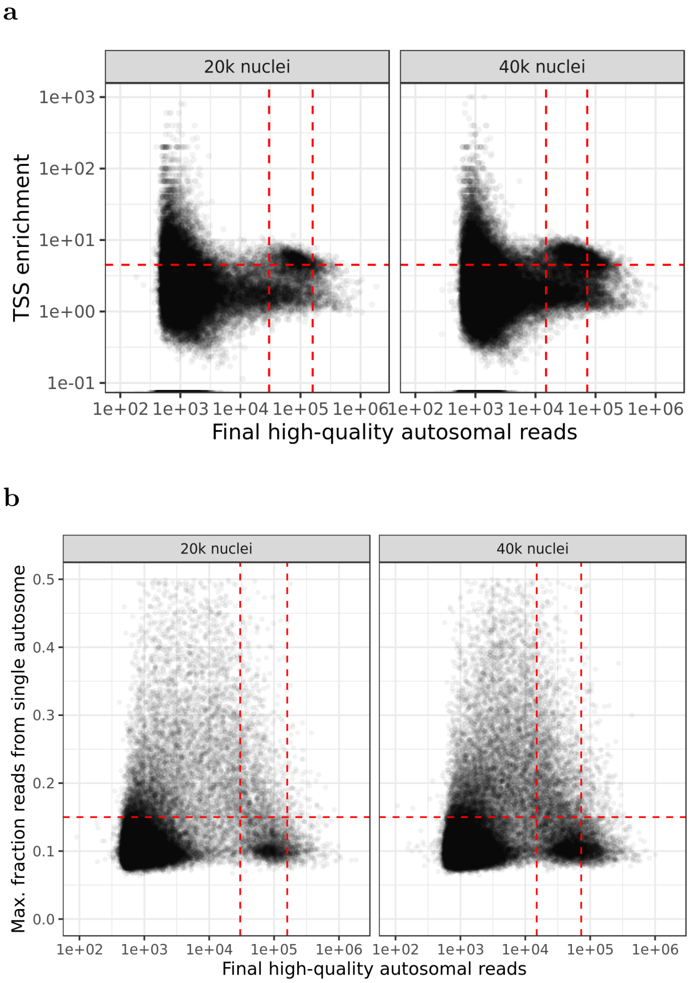


Figure 4.6 QC thresholds for the 20k and 40k nuclei input snATAC-seq libraries. (a) Dashed lines represent thresholds for minimum number of reads, maximum number of reads, and minimum TSS enrichment. (b) Dashed lines represent thresholds for minimum number of reads, maximum number of reads, and the maximum fraction of reads derived from a single autosome (imposed to filter out nuclei showing aberrant per-chromosome coverage).

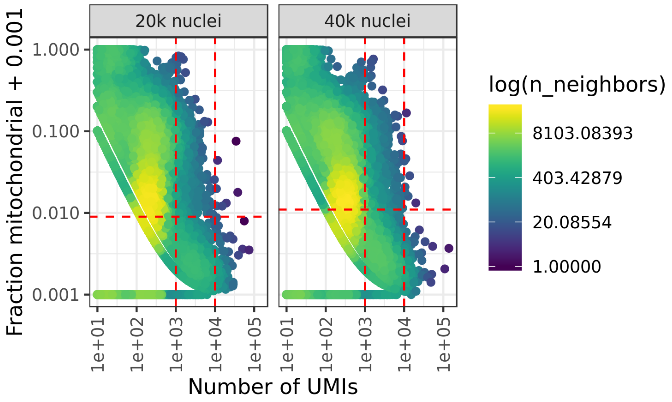


Figure 4.7 QC thresholds for the 20k and 40k nuclei input snRNA-seq libraries. Dashed lines represent thresholds for minimum number of UMIs, maximum number of UMIs, and maximum fraction of mitochondrial UMIs.

results (as measured by standard quality control metrics relative to bulk ATAC-seq data) and, for both snATAC-seq and snRNA-seq, increase the number of quality nuclei even after accounting for the increase in doublet rate. The aggregate gene expression/ATAC-seq signal profile was comparable between loading concentrations. One caveat to these conclusions is that the actual number of nuclei loaded into the well may differ from our estimated numbers, as debris in the nuclei preps may affect the accuracy of the nuclei counts.

4.3.3 Joint clustering of human and rat snATAC-seq and snRNA-seq identifies skeletal muscle cell types

To determine cell types present in skeletal muscle samples, we selected high-quality ATAC and RNA nuclei from the FANS/non-FANS libraries and the 20k/40k nuclei libraries generated above and performed joint clustering. snATAC-seq libraries that underwent FANS were excluded as they failed to provide quality data. We generated and included a snATAC-seq library containing a mix of human and rat nuclei (Fig. 4.8, 4.9). Information about the biological samples and post-QC nucleus summary statistics for each library is provided in Table 4.1. In total we obtained 24,866 human snRNA-seq (mean UMIs = 7,482), 5,053 human snATAC-seq (mean fragments

		Nuclei from sample		
library	**median/mean fragments**	**HSM1**	**HSM2**	**Rat**
Human/rat mix (ATAC)	49484/55905	887	0	612
no FANS, rep. 1 (ATAC)	75315/80666	670	0	0
no FANS, rep. 2 (ATAC)	53694/56209	606	0	0
no FANS, rep. 1 (RNA)	8610/9635	2004	0	0
FANS, rep. 1 (RNA)	8264/8799	7715	0	0
no FANS, rep. 2 (RNA)	8512/9770	2078	0	0
FANS, rep. 2 (RNA)	8172/8708	7118	0	0
20k nuclei (RNA)	2530/3030	1045	1073	0
20k nuclei (ATAC)	37110/38727	386	469	0
40k nuclei (RNA)	2190/2654	1913	1926	0
40k nuclei (ATAC)	20390/20991	955	1080	0

Table 4.1: Per-library fragment counts and per-sample nucleus counts for jointly clustered libraries

= 31,500), and 612 rat snATAC-seq (mean fragments = 60,874) nuclei. We used integrative non-negative matrix factorization (iNMF) as implemented in the LIGER (linked inference of genomic experimental relationships) software package [300] to perform joint clustering on snRNA-seq and snATAC-seq nuclei and identified seven cell type clusters (Fig. 4.10A). Nuclei from different modalities, species, and libraries integrated well, indicating that clustering was not driven by technical factors (Fig. 4.10B).

We used marker genes to assign cell types to each cluster (Table 4.2) and found clear concordance between human RNA-seq and ATAC-seq (Fig. 4.10C, D). We found marker gene accessibility in the rat ATAC-seq data to be largely consistent with the human data, though examination of the myosin heavy chain genes, often used to distinguish between different muscle fiber types, indicated that a considerable number of rat type II muscle fiber nuclei were likely present in the type I muscle fiber cluster (the opposite did not seem to occur; i.e., the type II muscle fiber cluster appeared to be relatively free of rat type I muscle fiber nuclei; Fig. 4.11). This mixing of some rat muscle fiber nuclei is a limitation of our data; because only 612 of 30,531 (2.0%) of

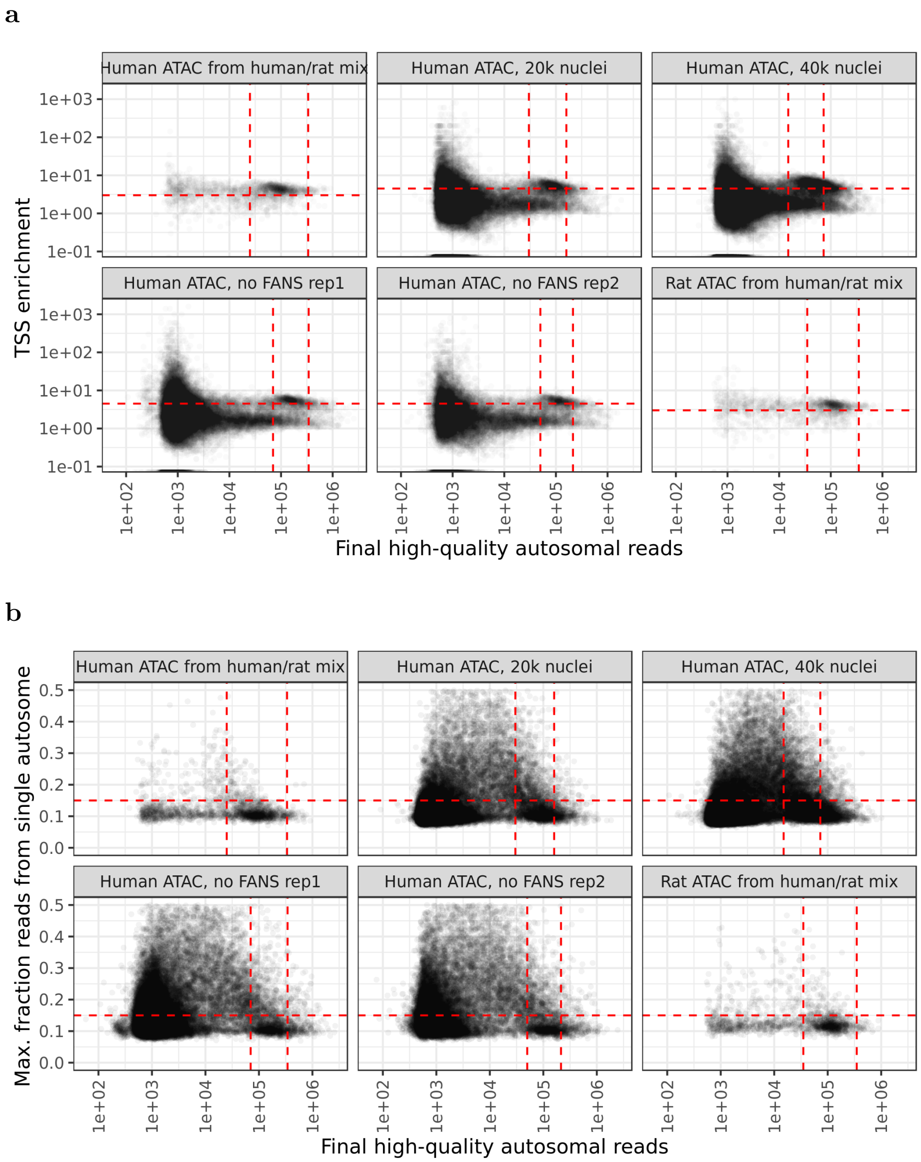


Figure 4.8 QC thresholds for all snATAC-seq libraries used in cell type clustering and downstream analyses. (a) Dashed lines represent thresholds for minimum number of reads, maximum number of reads, and minimum TSS enrichment. (b) Dashed lines represent thresholds for minimum number of reads, maximum number of reads, and the maximum fraction of reads derived from a single autosome (imposed to filter out nuclei showing aberrant per-chromosome coverage).

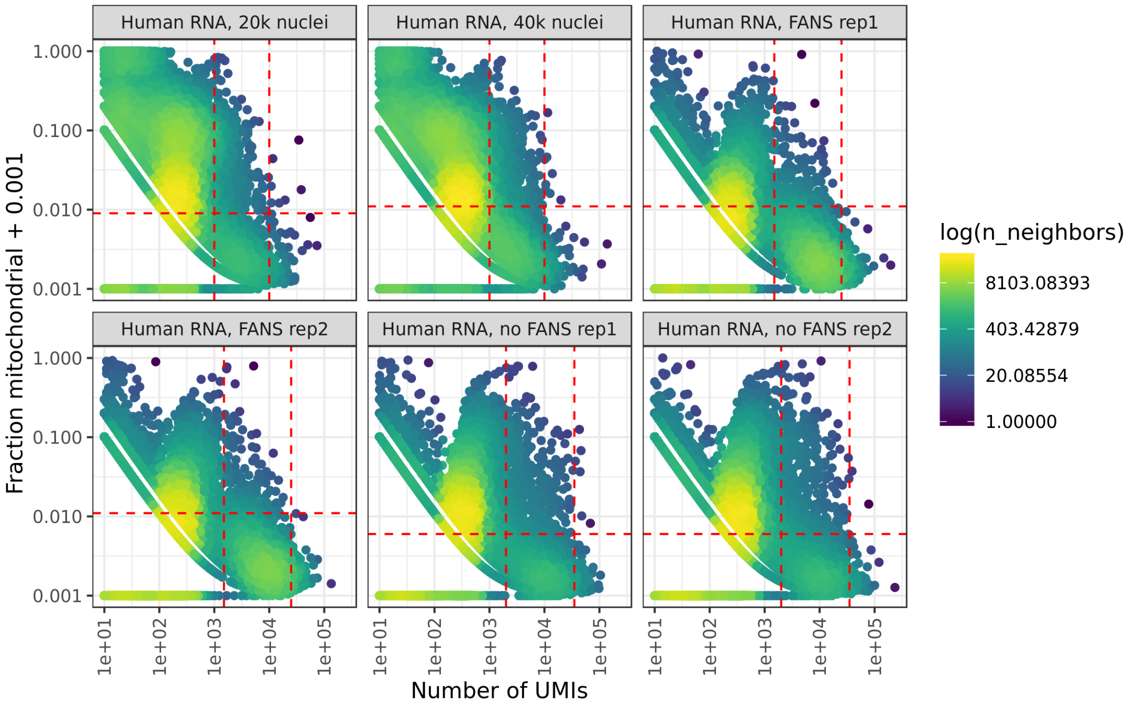


Figure 4.9 QC thresholds for all snRNA-seq libraries used in cell type clustering and downstream analyses. Dashed lines represent thresholds for minimum number of UMIs, maximum number of UMIs, and maximum fraction of mitochondrial UMIs.

all nuclei come from rat, the human data drive the clustering. As expected the vast majority of our nuclei (90.4%) were muscle fiber nuclei (Fig. 4.10E).

We sought to independently assess cluster identity without relying on marker gene patterns and therefore focused on cluster-level TSS-distal ATAC-seq peaks, many of which would not be taken into account when assigning cell types using marker genes. We developed a logistic regression approach to score the similarity between these

Gene	Cell type	References
MYH1	Type II muscle fibers	[255, 275]
MYH7	Type I muscle fibers	[255, 275]
PDGFRA	Mesenchymal stem cells	[69, 283]
VWF	Endothelial cells	[249, 312]
MYH11	Smooth muscle	[32, 88]
CD163	Immune (macrophages/monocytes)	[146, 137]
PAX7	Satellite cells	[235, 260]

Table 4.2: Marker genes used for cell type assignments

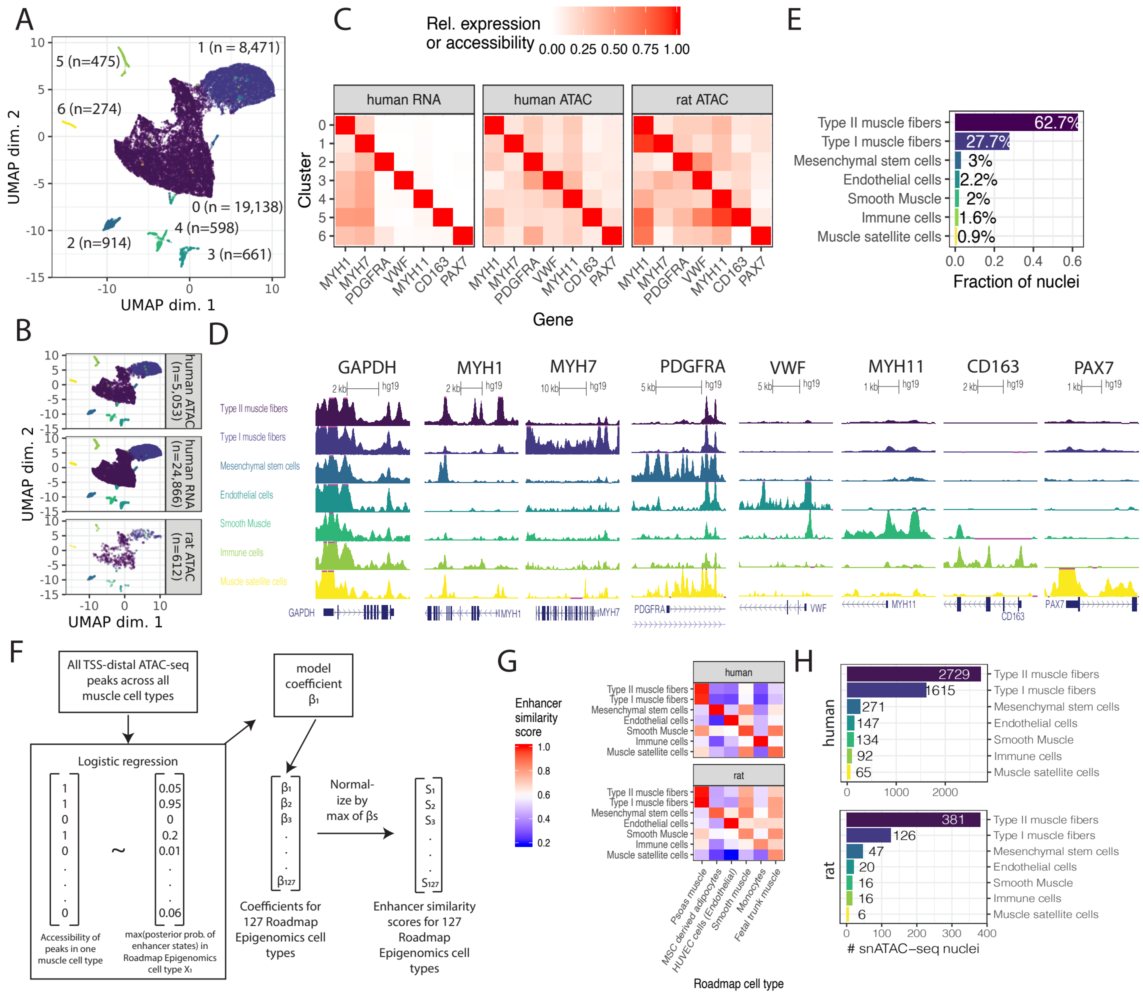


Figure 4.10 Joint clustering of snRNA-seq and snATAC-seq nuclei and cell type determination. (A) UMAP after clustering human snATAC-seq, human snRNA-seq, and rat snATAC-seq nuclei with LIGER. (B) UMAP faceted by species and modality. (C) Gene expression (snRNA-seq) or accessibility (snATAC-seq; gene promoter + gene body) of marker genes. Values are column-normalized. (D) ATAC-seq signal for human snATAC-seq nuclei in each cluster. All tracks are normalized to 1M reads. (E) Fraction of nuclei, across both species and modalities, assigned to each cell type. (F) Logistic regression-based approach to score similarity between TSS-distal ATAC-seq peaks (> 5 kb from TSS) and Roadmap Epigenomics enhancer states. (G) Similarity of snATAC-seq peak calls for each cell type and species to Roadmap Epigenomics chromHMM enhancer states based on the logistic regression procedure outlined in (F). The Roadmap Epigenomics cell type names have been adjusted slightly for clarity and the sake of space. The full names and the identifiers from the Roadmap Epigenomics paper are: Psoas muscle (E100), Mesenchymal Stem Cell Derived Adipocyte Cultured Cells (E023), HUVEC Umbilical Vein Endothelial Primary Cells (E122), Stomach Smooth Muscle (E111), Primary monocytes from peripheral blood (E029), and Fetal Muscle Trunk (E089). (H) Nucleus counts per species for snATAC-seq data.

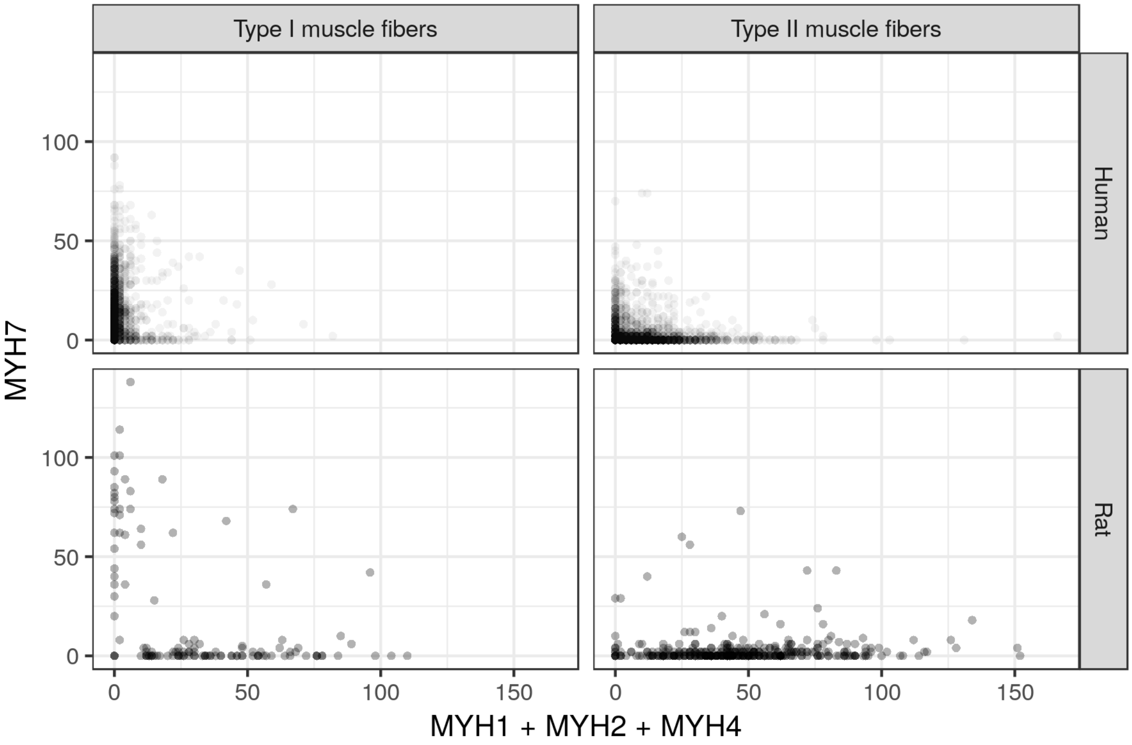


Figure 4.11 snATAC-seq read counts (gene promoter + gene body) derived from the Type II muscle fiber myosin heavy chain genes (*MYH1*, *MYH2*, *MYH4*) or the Type I muscle fiber myosin heavy chain gene (*MYH7*) for human and rat nuclei. Each point represents a single nucleus. Type I muscle fibers/Type II muscle fibers headers represent the cluster to which each nucleus was assigned.

Cell type	Human ATAC	Human RNA	Rat ATAC
Type II muscle fibers	2729	16028	381
Type I muscle fibers	1615	6730	126
Mesenchymal stem cells	271	596	47
Endothelial cells	147	494	20
Smooth Muscle	134	448	16
Immune cells	92	367	16
Muscle satellite cells	65	203	6

Table 4.3: Cell type nucleus counts by species and modality

peaks and enhancer chromatin states from 127 Roadmap Epigenomics cell types (Fig. 4.10F) [243]. We found concordance with the marker gene-based cell type assignment approach (Fig. 4.10G). Remarkably this approach worked relatively well in assigning rat nuclei, despite the fact that the number of nuclei per cluster for rat ranged between six and twenty for the smallest four cell types (Table 4.3; Fig. 4.10H).

The majority of the nuclei were assigned as type I or type II muscle fibers. Genes previously discovered to be preferentially expressed in type I vs. type II muscle fibers [246] were usually similarly preferentially expressed in our snRNA-seq data (Fig. 4.12), validating the quality of the data and accuracy of muscle fiber type assignments.

4.3.4 Integration of cell-type-specific ATAC-seq peaks with UK Biobank GWAS reveals cell type roles in complex phenotypes

Genetic variants associated with complex traits and disease are frequently located in non-coding regions of the genome [179, 208, 253]. Variants associated with a given complex trait are expected to be enriched specifically in non-coding regulatory elements of the trait-relevant cell types; for example, T2D-associated genetic variants are enriched in regulatory elements specific to pancreatic islets and beta cells [71, 79, 172, 208, 209, 224, 228, 281, 286, 287], and variants associated with autoimmune disorders are enriched in immune cell-specific regulatory elements [71]. Variant

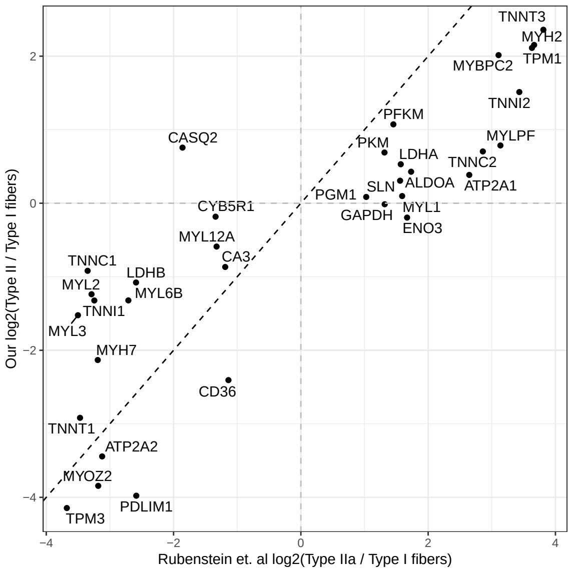


Figure 4.12 Log2(fold change) for Type II vs Type I muscle fiber gene expression, showing the top differentially expressed genes from Rubenstein et al. (Rubenstein et al. Table S4).

enrichment in cell-specific regulatory elements can therefore be used to determine which cell types are relevant to a given trait or disease. Variants in high linkage disequilibrium (LD) with trait-influencing SNPs are often statistically associated with the trait as well, making it difficult to infer the causal SNP through statistical association alone. Epigenomic data, such as chromatin accessibility in trait-relevant cell types, can be used to nominate causal genetic variants under the assumption that non-coding SNPs in accessible regions of the genome are more likely to be causally related to a trait than non-coding SNPs in inaccessible regions.

To explore the relationship between complex traits and the cell types present in our data, as well as demonstrate the value of our muscle cell type chromatin data in narrowing the post-GWAS search space, we used LD score regression (LDSC) [71, 80] to perform a partitioned heritability analysis using GWAS of 404 heritable traits from the UK Biobank [270] (`http://www.nealelab.is/uk-biobank/`) and our muscle cell type open chromatin regions [71, 80]. Results for all traits in which at least one of our cell types showed significant ($p < 0.05$) enrichment after Benjamini-Yekutieli correction are displayed in Fig. 4.13A. Due to the heavy multiple testing correction burden, relatively few traits meet this threshold. However, we observed immune cell abundance traits show enrichment for the immune cell cluster, and diastolic blood pressure GWAS SNPs are enriched in smooth muscle ATAC-seq peaks. In addition, we see that several skeletal trait GWAS SNPs are enriched in mesenchymal stem cell peaks. Previous work has shown a central role of bone mesenchymal stem cells in osteoblast development [216, 309]. In addition, SNPs for several corneal traits are also enriched in mesenchymal stem cell peaks, consistent with previously observed enrichment of corneal thickness GWAS SNPs in mesenchymal stem cell/connective tissue cell annotations [105]. Results using rat peaks projected into human coordinates largely mirror the human mesenchymal stem cell enrichment findings (Fig. 4.14).

One muscle-related trait included in the UK Biobank is creatinine level. In humans most serum creatinine is produced by skeletal muscle and is filtered by the kidneys [116]. Creatinine levels are commonly used as a biomarker for kidney function but correlate with muscle mass and have been used to score sarcopenia [12, 117, 154]. In our enrichment analysis, the cell type with the highest LDSC coefficient Z-score was type II muscle fibers (z-score = 2.5; Fig. 4.13B).

Integrating the ATAC-seq results with the GWAS summary statistics can help nominate causal SNPs. One example is the *C17orf67* locus in the creatinine GWAS (Fig. 4.13C). The lead SNP at this locus (rs227727; p = 5.38e-18) lies in an intergenic region 92 kb from *C17orf67* and 104 kb from *NOG*. This SNP is in an ATAC-seq peak in several muscle cell types, though the signal is largest in type II muscle fibers (Fig. 4.13D). The peak corresponds to an enhancer chromatin state in muscle, amongst other cell types [243]. We used the Probabilistic Identification of Causal SNPs (PICS) tool [70] to estimate the probability that nearby SNPs were causal given the pattern of linkage disequilibrium at the locus. PICS assigned the

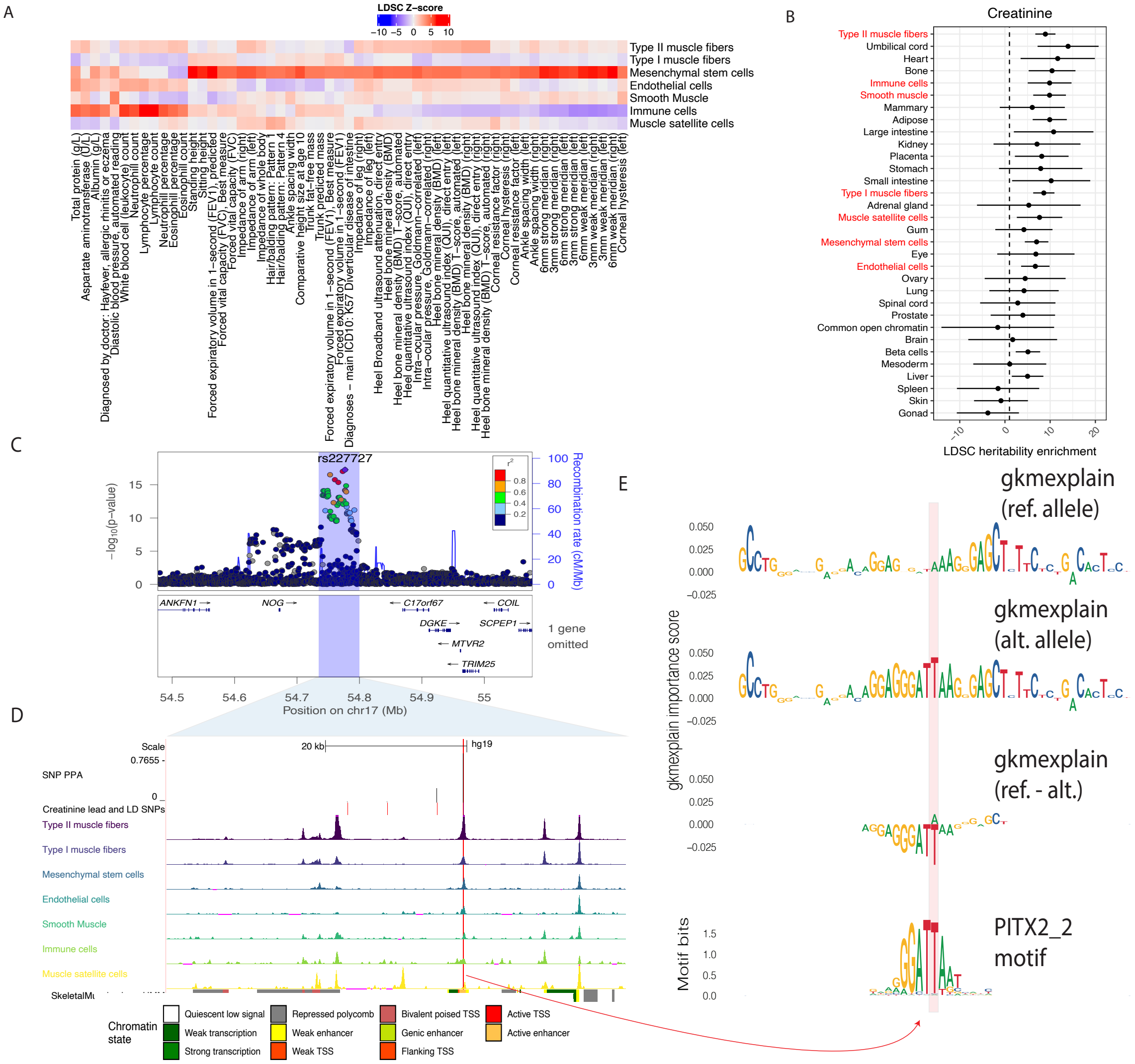


Figure 4.13 Integration of UK Biobank GWAS with cell type snATAC-seq peaks. (A) UK Biobank LDSC partitioned heritability results for traits for which one of the muscle cell types was significant after Benjamini-Yekutieli correction. (B) LDSC partitioned heritability results for creatinine (UK Biobank trait 30700). Red y-axis labels refer to the muscle snATAC-seq cell type annotations. (C) Locuszoom plot [221] for *C17orf67* locus in the UK Biobank creatinine GWAS. (D) ATAC-seq signal in the region highlighted in (C). All tracks are normalized to 1M reads. SNPs shown have LD ≥ 0.8 with the lead SNP based on the European samples in 1000 Genomes Phase 3 (Version 5) [1]. (E). gkmexplain importance scores for the ref and alt allele-containing sequences (top two rows), and the difference between the ref and alt allele importance scores (third row), which resembles the PITX2_2 motif predicted to be disrupted by the A allele (bottom row).

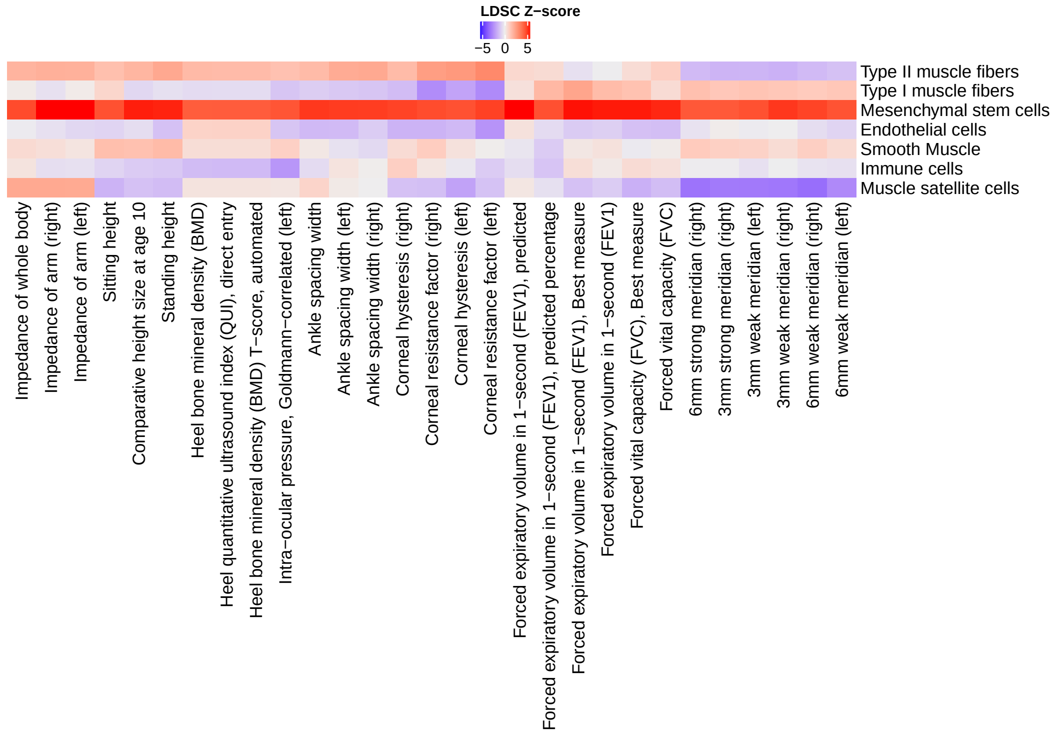


Figure 4.14 UK Biobank LDSC partitioned heritability results for traits for which one of the muscle cell types was significant after Benjamini-Yekutieli correction (rat).

index SNP, rs227727, a probability of 0.766 of being the causal SNP. A tightly linked SNP, rs227731 ($R^2 = 0.99$), had a probability of 0.221; no other SNPs had probability greater than 0.01. SNP rs227731 is not in an ATAC-seq peak in any of the muscle cell types we identified, suggesting that the index SNP, rs227727, is indeed the causal SNP. A previous study found that the A allele of rs227727 was associated with higher activity in an allelic luciferase assay in both human fetal oral epithelial cells (GMSM-K) and murine osteoblastic cells (MC3T3) [153]. To predict allelic effects at this SNP in type II muscle fibers, we trained a gapped-kmer support vector machine model (gkm-SVM) [82, 150] to detect kmers associated with increased or decreased chromatin accessibility using the top ATAC-seq peaks for each of our cell types and then ran deltaSVM [151] to predict this SNP's effect on chromatin accessibility. DeltaSVM predicts a SNP's effect by comparing the gkm-SVM inferred kmer weights for kmers created by the reference vs the alt allele; we transformed the deltaSVM score to a z-score based on the distribution of the predicted impacts of all autosomal 1000 Genomes SNPs [1]. The type II muscle fiber deltaSVM z-score for this SNP was 0.73 (directionally favoring the alt allele, T, having higher chromatin accessibility, although the z-score is not statistically significant). We also attempted to interpret how each allele of the SNP affects the gkm-SVM model's score for the sequence using the gkmexplain software package, which scores the importance of each base in a sequence to the gkm-SVM model score for the sequence [265]. We ran gkmexplain on the sequence surrounding the SNP in the presence of either the reference or the alternative allele and compared the results (Fig. 4.13E). The change in the gkmexplain importance scores in the presence of the reference vs alternative allele resembled several known homeodomain TF motifs predicted to be disrupted by the reference allele such as that of PITX2, suggesting that the alternate allele may have directionally (non-significant) greater predicted chromatin accessibility because it is a better match to these homeodomain TF motifs (Fig. 4.13E) [125]. We note, however, that the deltaSVM z-score of the SNP as well as the gkmexplain importance scores of the SNP and surrounding nucleotides are of low magnitude, suggesting that the reference allele may reduce the binding of PITX2 or another homeodomain TF without significantly affecting local chromatin accessibility. Biologically, the nearby *NOG* gene is a particularly compelling candidate target gene of this regulatory element, as its product (noggin) regulates BMP signaling and is involved in muscle growth and maintenance [47, 73, 251, 250, 284, 293]. Integrated with the GWAS summary statistics and these additional resources, our ATAC-seq data adds to existing evidence that SNP rs227727 alters the activity of a gene regulatory element and is a prime candidate to impact creatinine levels.

4.3.5 Integration of cell type-specific ATAC-seq peaks with T2D GWAS credible sets nominates causal SNPs and cell types

It is well-established that T2D GWAS SNPs overlap pancreatic islet/beta cell enhancers [79, 172, 208, 228, 286]; however, some SNPs may act through other T2D-relevant tissues, such as muscle, adipose, or liver. We therefore used LDSC to perform a partitioned heritability analysis for T2D-associated SNPs [172] in each of the muscle cell types as well as in beta cell ATAC-seq peaks, adipose ATAC-seq peaks, and liver DNaseI hypersensitive sites (see Methods) (Figs. 4.15A, 4.16A). When modeling each cell type separately (adjusting for the cell type-agnostic LDSC baseline annotations and common open chromatin regions), we found significant enrichment (after Bonferroni correction for 40 tests) in type II muscle fibers and beta cells, though when modeling all cell types in a single joint model only beta cell open chromatin regions showed significant enrichment (Fig. 4.16A). We performed a similar analysis on GWAS SNPs for a T2D-related trait, fasting insulin (Figs. 4A, 4.16A) [175]. For fasting insulin, we found significant enrichment in mesenchymal stem cells, immune cells, and bulk adipose when modeling each cell type individually, but only adipose showed significant enrichment when modeling all cell types jointly. For fasting insulin, we note that the small sample size of that GWAS means the analysis was likely underpowered, leaving open the possibility that other cell types will show significant enrichment when GWAS with larger sample sizes are available. We also note that the adipose open chromatin regions are derived from bulk tissue open chromatin profiling; it is therefore possible that at least some of the signal from adipose is being driven by cell types shared between our muscle samples and adipose tissue, such as mesenchymal stem cells. This is an area for further exploration when single-cell/single-nucleus data from adipose is available.

We performed similar GWAS enrichments using the rat muscle cell type peaks projected into human coordinates (Fig. 4.15A, 4.16B). For T2D we found muscle fiber types and mesenchymal stem cells were significantly enriched after Bonferroni correction, but as with human muscle cell types these enrichments did not persist in a joint model with all cell types (Fig 4.16B). For fasting insulin no rat muscle cell types showed enrichment after Bonferroni correction.

While none of our cell types showed significant enrichment in 10-cell-type models after Bonferroni correction, it is still possible that some T2D GWAS loci act through muscle cell types or cell types shared between muscle and other tissues such as adipose. There are a substantial number of T2D GWAS credible sets that show no overlap with pancreatic islet functional annotations [172]. We therefore overlapped 380 previously-published T2D GWAS signals with 99% genetic credible set SNPs [172] with our snATAC-seq peaks to nominate SNPs that may be acting through the muscle cell types, including those that are expected to be shared with adipose.

One locus highlighted by our data is the *ITPR2* locus on chromosome 12 (Fig. 4.15B). This locus contains 22 credible set SNPs, none with a particularly high posterior probability of association (PPA) in the T2D genetic fine-mapping (maximum across all credible set SNPs = 0.06). Only one SNP (rs7132434; PPA = 0.042) overlaps any of our muscle cell type peak calls (Fig. 4.15C). This SNP is in a large mesenchymal stem cell ATAC-seq peak, and also overlaps peak calls in smooth muscle and blood, though the chromatin accessibility signal in those cell types is lower in our data. The SNP also overlaps a peak call in a subset of adipose and islet samples (Fig. 4.17). We found that this SNP had a large deltaSVM z-score in several of the muscle cell types (absolute z-score = 2.88 in mesenchymal stem cells; the T2D risk allele, A, is predicted to result in greater chromatin accessibility). We ran gkmexplain on the sequence surrounding the SNP and found the gkmexplain importance scores for the sequence in the presence of the risk allele resembled an AP-1 motif (Fig. 4.15D) [125]. A literature search revealed that the element underlying this SNP has been validated for enhancer activity using a luciferase assay (in a renal cancel cell line, 786-O cells)

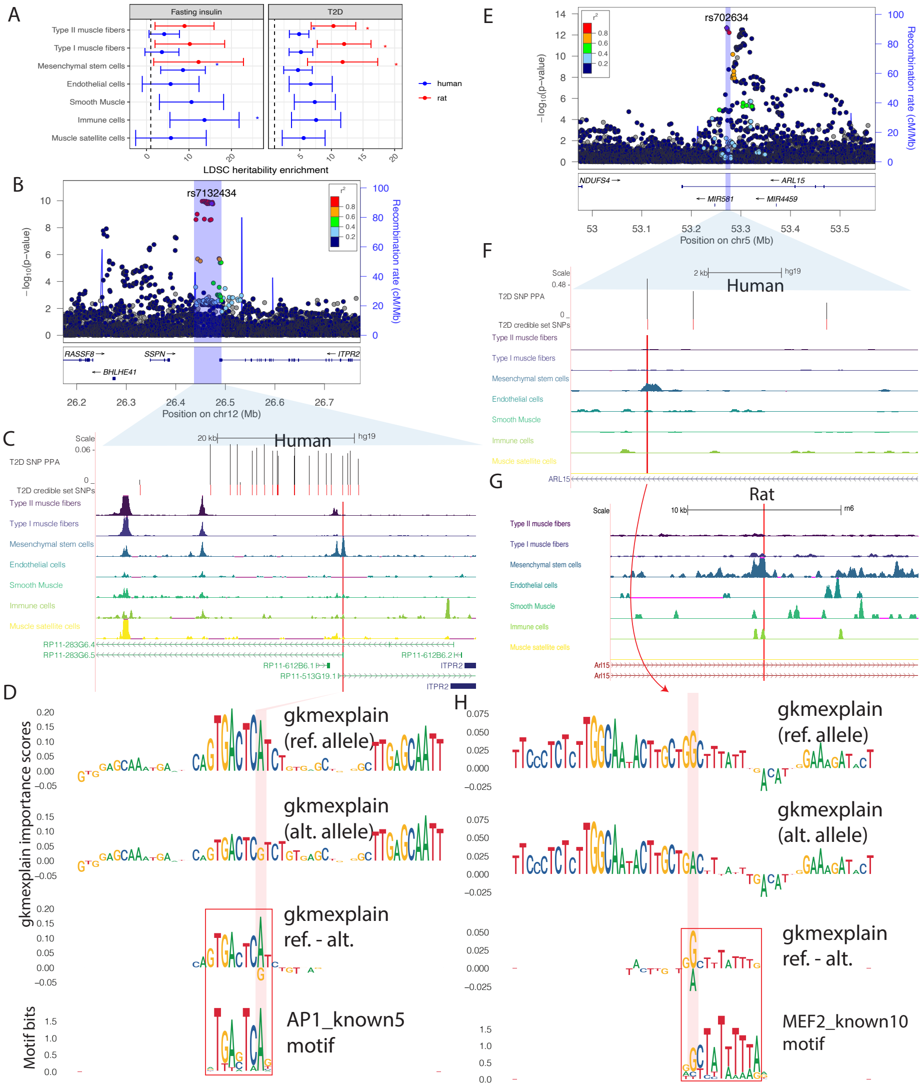


Figure 4.15 Integration of T2D and related trait GWAS with cell type snATAC-seq peaks. (A) LDSC partitioned heritability results for T2D and Fasting insulin GWAS, using human peak calls. For each of the cell types, one model was run adjusting for cell type-agnostic annotations from the LDSC baseline model and common open chromatin regions. Asterisks represents Bonferroni significance (p < 0.05 after adjusting for 40 tests). (B) locuszoom plot [221] for *ITPR2* locus in the T2D data. (C) T2D credible set near the *ITPR2* gene, consisting of 22 SNPs. One SNP (highlighted in red) overlaps a peak call in any of the muscle cell types. (D) gkmexplain importance scores for the ref and alt allele (top two rows) and the difference between the ref and alt importance scores (third row); the G allele disrupts an AP1 motif (bottom row). (E). locuszoom plot for *ARL15* locus in the T2D data. (F). T2D credible set SNPs near the *ARL15* gene. One of the SNPs overlaps a mesenchymal stem cell specific peak. (G). Projecting the SNP highlighted in (F) into the rat genome (projected SNP position indicated by the red vertical line) shows the corresponding region has open chromatin in rat mesenchymal stem cells. (H). gkmexplain importance scores for the ref and alt alleles (top two rows), the difference between them (third row), and a MEF2 motif disrupted by the SNP.

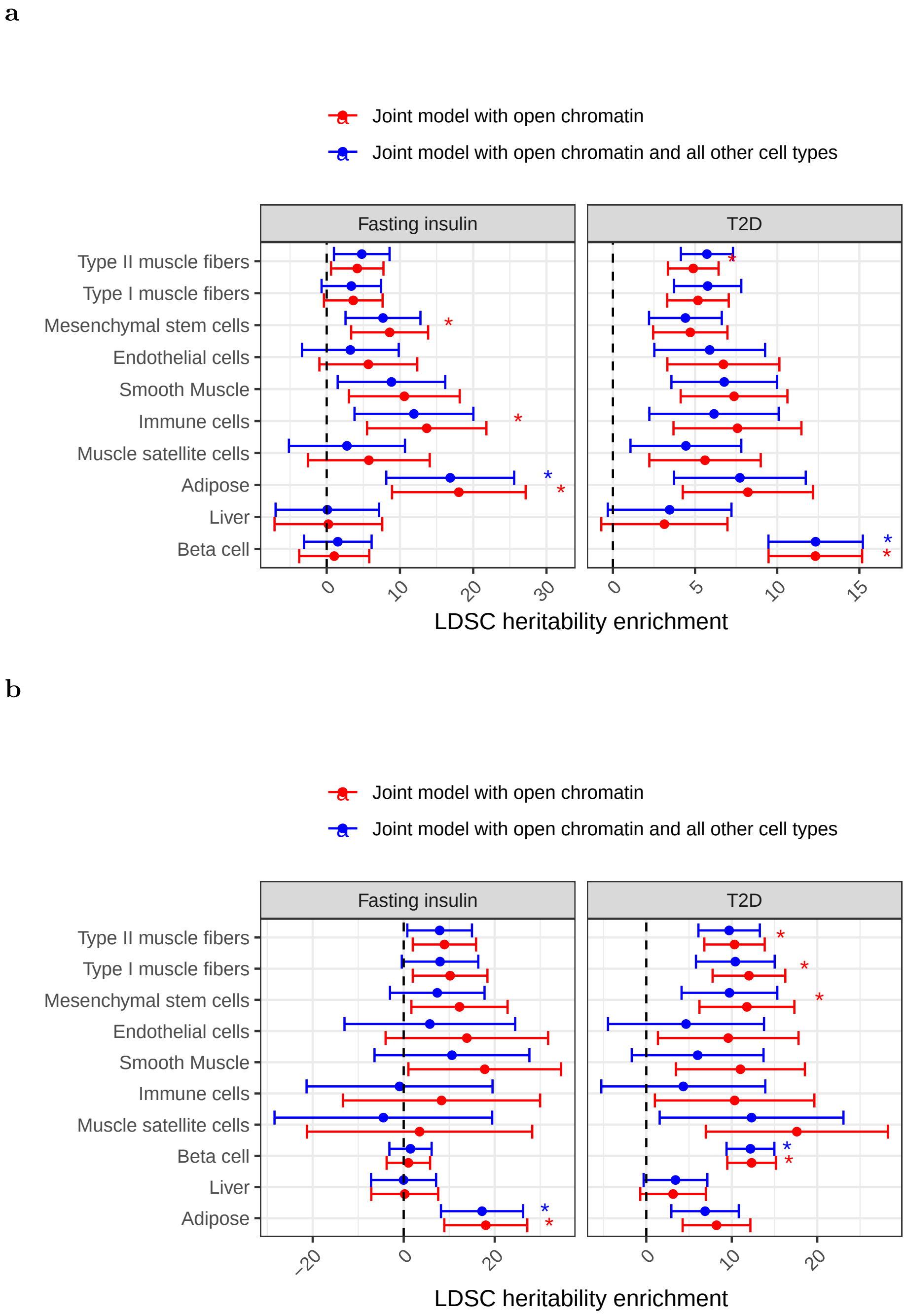


Figure 4.16 (A) LDSC partitioned heritability results for T2D and Fasting insulin GWAS, using human peak calls. Results are shown for pancreatic beta cell, adipose, and liver open chromatin regions as well. First, for each of the ten cell types, one model was run adjusting for cell type-agnostic annotations from the LDSC baseline model and common open chromatin regions (this is the joint model with open chromatin). Then, a single model containing those same annotations and all ten cell types was run (this is the joint model with open chromatin and all other cell types). Asterisk represents Bonferroni significance ($p < 0.05$ after adjusting for two traits, ten cell types, and two models per cell type = 40 tests). (B) Same as (A), but using the rat peak calls projected into human coordinates for the muscle cell types.

and the risk allele showed preferential binding of the AP-1 transcription factor in an EMSA assay in the same study and cell line [17], consistent with our findings. We note that this SNP is also a credible set SNP for waist-hip ratio [99, 162]. We therefore hypothesize that rs7132434 is the causal SNP at this locus, and that it may be acting through mesenchymal stem cells.

A second locus highlighted by our data is an intronic locus in the *ARL15* gene (Fig 4E). The T2D genetic fine-mapping narrowed the list of potentially causal SNPs at this locus to three (two other, larger genetic fine-mapping credible sets are also annotated to *ARL15*). SNPs in this credible set are statistically associated with fasting insulin [171], and more broadly variants in or near *ARL15* associate with metabolic traits including adiponectin, HDL cholesterol levels, and BMI [171, 241, 278], suggesting that the locus may affect T2D risk not through islets but through adipose or a related cell type. Interestingly, none of the SNPs overlap with any of ENCODE's 1.3 million candidate cis-regulatory elements [64, 65] or any of the approximately 3.6 million DNaseI hypersensitive sites (DHS) annotated in [184]; however, in our data we find that one of the SNPs (rs702634) is in the center of a mesenchymal stem cell specific ATAC-seq peak (Fig. 4.15F), and a mesenchymal stem cell peak is likewise present in the corresponding position in the rat genome (Fig. 4.15G), indicating that this is a regulatory element that has been conserved across species. The T2D genetic fine-mapping assigned this SNP a probability of 0.48 of being the causal SNP at this locus, higher than either of the other two SNPs (0.33 and 0.19, respectively). We examined publicly-available beta cell (n = 1), islet (n = 10) [228], and adipose (n = 3) [29] ATAC-seq data to see if hints of this peak are present in these T2D-relevant cell types. No convincing signal appears to be present in beta cell or islet data (a subset of islet samples do have a peak call overlapping the SNP, but visual inspection of the locus suggests this is spurious); a weak but significant and visually apparent increase in signal at that SNP is evident in adipose samples (Fig. 4.18). As mes-

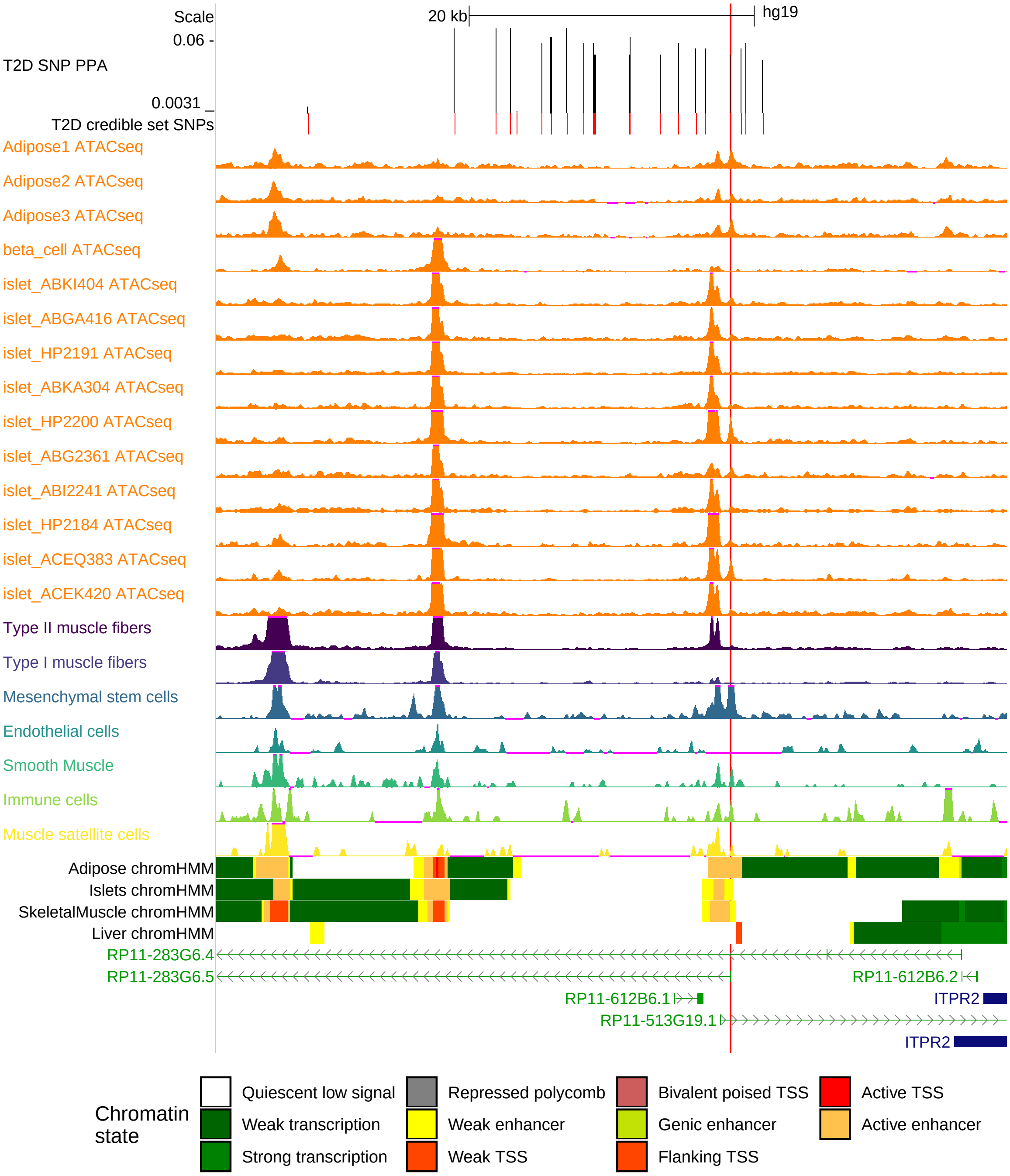


Figure 4.17 ATAC-seq signal in bulk adipose, bulk islet, single-nucleus pancreatic beta cell, or our muscle cell types at the *ITPR2* locus credible set. All tracks are normalized to 1M reads.

enchymal stem cells are one component of adipose tissue, it is possible that the weak signal in adipose is due to mesenchymal stem cell populations within adipose; this is one area for follow-up when adipose single-nucleus ATAC-seq data is available. The absolute deltaSVM z-score in mesenchymal stem cells for this SNP was 0.48, indicating it does not have a large impact on predicted chromatin accessibility; however, the risk allele is predicted to disrupt a MEF2 motif [90, 125], and we found the change in gkmexplain importance scores between the reference and alternative allele showed some resemblance to this motif (Fig. 4.15H). This data is consistent with a model in which rs702634 is the causal SNP and acts through mesenchymal stem cells.

4.4 Discussion

Here we present snATAC-seq and snRNA-seq for human skeletal muscle and snATAC-seq for rat skeletal muscle, which we use to map the transcriptomes and chromatin accessibility of cell types present in skeletal muscle samples. The cell types identified are consistent with known biology and with previous studies of human [246] and mouse [54, 83, 274] skeletal muscle tissue. However, our use of single-nucleus rather than single-cell techniques allows us to capture muscle fiber nuclei, cell types missing from previously published snRNA-seq datasets. To our knowledge this is the first published snATAC-seq dataset for human and rat skeletal muscle tissue. We therefore anticipate that this dataset will be useful in nominating causal GWAS SNPs and demonstrate this by integrating the data with UK Biobank and previously published T2D GWAS credible sets, highlighting potentially causal SNPs at the *C17orf67*, *ARL15*, and *ITPR2* loci.

Additionally, we explore the effect of two technical parameters on snRNA-seq and snATAC-seq results. First, we find that FANS substantially alters snATAC-seq results. Though the stereotypical ATAC-seq fragment length distribution is observed, signal-to-noise (as measured by TSS enrichment and fraction of reads in peaks, as

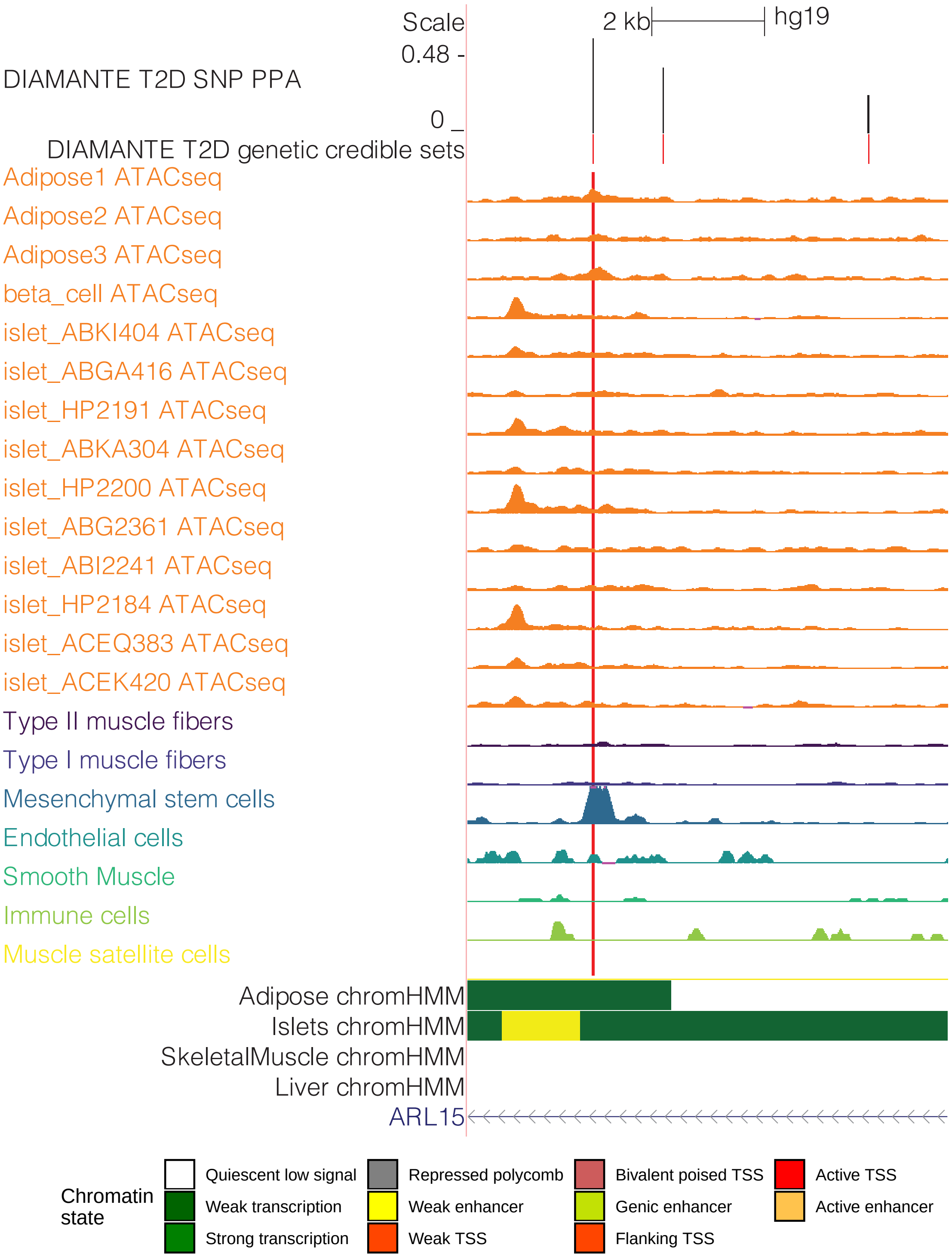


Figure 4.18 ATAC-seq signal in bulk adipose, bulk islet, single-nucleus pancreatic beta cell, or our muscle cell types at the *ARL15* locus. All tracks are normalized to 1M reads.

well as by visual inspection) appears to decrease substantially relative to non-FANS libraries. We note that the effect of FANS (nucleus sorting) may differ from that of FACS (cell sorting). snRNA-seq results appear to be substantially less sensitive to FANS – the pseudobulk gene expression from FANS libraries correlates strongly with that from non-FANS libraries – suggesting that chromatin is more sensitive to FANS than is RNA. We also observed higher nucleus yield in our FANS snRNA-seq libraries than our non-FANS libraries. There are several potential explanations for this. One is that the nuclei counting step that necessarily precedes loading of the 10X platform may be sensitive to debris. If greater amounts of debris are observed in non-FANS libraries, nucleus concentration may be systematically overestimated in non-FANS libraries, resulting in more nuclei actually being loaded onto the 10X platform from FANS libraries. While not mutually exclusive, FANS may also decrease the amount of debris being loaded into the 10X platform, and thereby improve nucleus capture.

We found snATAC-seq and snRNA-seq results were remarkably consistent at different loading concentrations. One clear caveat is that this may change as the loading concentration is further reduced or increased. It is also important to note that the actual number of nuclei loaded may differ from the estimated 20k or 40k nuclei. As discussed above, it is possible that debris in the input preparation makes nucleus counting less accurate, in which case our cited values may not reflect the true values. However, because the same nuclear preparation was used as input for the 20k and 40k nuclei libraries, the two-fold difference in loading concentration should be reliable, even if the absolute values are skewed.

The GWAS enrichments presented here will be one interesting area to follow up on as more snATAC-seq data is published. Interpretation of the results is complicated by the fact that many tissues share cell types. For example, mesenchymal stem cell-like populations exist in many tissues besides muscle, such as adipose tissue and bone marrow. Taking the fasting insulin enrichments as an example, we found that the enrichment of GWAS SNPs in muscle cell type ATAC-seq peaks disappeared when adipose tissue was included in the enrichment model. However, it is possible that the adipose enrichment is being driven in part by mesenchymal stem cell populations within adipose itself. Direct comparison of snATAC-seq and snRNA-seq profiles from mesenchymal stem cells from a wider array of tissues will help tease apart commonalities and tissue-specific differences in this interesting population.

4.5 Methods

4.5.1 snATAC-seq and snRNA-seq, FANS vs no FANS experiment

Three separate pieces of tissue were cut from a single human skeletal muscle sample (weighing 60mg, 50mg and 50mg; sample HSM1). Nuclei were isolated using a modified version of the ENCODE protocol [65], customized from Step 5 onwards to accommodate FANS (Fluorescence assisted nuclei sorting). In step 5, the nuclei were resuspended in 700 μL of Sort buffer (1% BSA, 1mM EDTA in PBS) and filtered through a 30 μm filter. Three different nuclei isolations were performed and the nuclei suspended in sort buffer were mixed, pooled together and divided into two groups, one with FANS and one without FANS. FANS nuclei were sorted according to the previously published FANS protocol using DRAQ7 [219]. DRAQ7 (0.3mM from Cell Signaling Technology) was added to the FANS nuclei suspension, at 100 fold dilution to get a final concentration of 3 μM. Nuclei were gently mixed and incubated for 10 minutes on ice. Nuclei were analyzed in the presence of DRAQ7 and sorted for high DRAQ7 positive signal using Beckman Coulter's Astrios MoFlo. We followed the gating strategy outlined in the FANS protocol [219]. The sorted nuclei were collected in a recovery buffer (5% BSA in PBS). The nuclei with and without FANS were spun at 1000g for 15 min at 4°C. The nuclei were resuspended in 100 μL of 1X diluted nuclei buffer and counted in the Countess II FL Automated Cell Counter. The appropriate amount of nuclei were split for snRNA-seq and spun down at 500g for 10 min at 4°C and resuspended in RNA nuclei buffer (1% BSA+PBS in 0.2U RNAse inhibitor). The nuclei at appropriate concentration for snATAC-seq and snRNA-seq were submitted to the Advanced Genomics core for all the snATAC-seq and snRNA-seq processing on the 10X Genomics Chromium platform (v. 3.1 chemistry for snRNA-seq). For each modality nuclei were loaded at 15.4K nuclei/well.

4.5.2 snATAC-seq and snRNA-seq, loading 20k or 40k nuclei

Two pieces of tissue (weighing 85.3 mg and 85.8 mg) were cut from one human skeletal muscle sample HSM1 and two tissue pieces (weighing 95.9 mg and 92.6 mg) were cut from a second human skeletal muscle sample, HSM2. Each of the samples was cut on dry ice using a frozen scalpel to prevent thawing. The samples were pulverized using an automated dry cryo pulverizer (Covaris 500001). We isolated nuclei using a customized protocol derived from the previously published ENCODE protocol [65]. All four pulverized tissues pieces were mixed and redistributed to perform four different nuclei isolations. The desired concentration of nuclei was achieved by resuspending the appropriate number of nuclei in 1X diluted nuclei buffer for snATAC-seq and RNA nuclei buffer (1% BSA in PBS containing 0.2U/μL of RNAse inhibitor) for snRNA-seq. The nuclei at appropriate concentration for snATAC-seq and snRNA-seq were submitted to the Advanced Genomics core for all the snATAC-seq and snRNA-seq processing on the 10X Genomics Chromium platform (v. 3.1 chemistry for snRNA-seq). For each modality nuclei were loaded at two different concentrations, 20K nuclei/well and 40K nuclei/well.

4.5.3 snATAC-seq, human and rat mixed library

Tissue from human (49mg of pulverized human skeletal muscle; sample HSM1) and rat (45mg of pulverized gastrocnemius samples) were used in this single nuclei

ATAC experiment. We used the previously published ENCODE protocol [65] to isolate nuclei, which is compatible with both snATAC-seq and snRNA-seq. After isolating nuclei from each sample (species) individually, the nuclei were mixed in equal proportions. The desired concentration of nuclei was achieved by resuspending the appropriate number of nuclei in 1X diluted nuclei buffer for snATAC-seq. The nuclei at the appropriate concentration for snATAC were submitted to the Advanced Genomics core for all the snATAC-seq processing on the 10X Genomics Chromium platform. 15.4K nuclei were loaded into a single well.

4.5.4 Bulk ATAC-seq

2 tissue pieces weighing 99.4 mg and 80.7 mg were cut from one human skeletal muscle sample (HSM1) and 2 pieces weighing 67.6 mg and 103.5 mg were cut from a second human skeletal muscle sample (HSM2). Each of the samples was cut on dry ice using frozen scalpel to prevent thawing. The samples were pulverized using an automated dry cryo pulverizer (Covaris 500001). After nuclei isolation, the nuclei were counted in Countess II FL Automated Cell Counter, and the appropriate volume of the suspension for 50K nuclei was spun down and used for the downstream transposition reaction (a modified version of the ENCODE protocol [65]).

4.5.5 Processing of muscle bulk ATAC-seq data

Adapters were trimmed using cta (v. 0.1.2; `https://github.com/ParkerLab/cta`). Reads were mapped to hg19 using bwa mem (-I 200,200,5000 -M; v. 0.7.15-r1140) [156]. Duplicates were marked using picard MarkDuplicates (v. 2.21.3; `https://broadinstitute.github.io/picard/`). We used samtools to filter to high-quality, properly-paired autosomal read pairs (-f 3 -F 4 -F 8 -F 256 -F 1024 -F 2048 -q 30; v. 1.9 using htslib v. 1.9) [157]. To call peaks, we used bedtools bamtobed [226] to convert to a bed file (v. 2.27.1) and then used that file as input to MACS2 callpeak

(–nomodel –shift -100 –seed 762873 –extsize 200 –broad –keep-dup all –SPMR; v. 2.1.1.20160309) [317]. To visualize the signal, we converted the bedgraph files output by MACS2 to bigwig files using bedGraphToBigWig (v. 4) [123].

4.5.6 Processing of snATAC-seq data

Adapters were trimmed using cta. We used a custom python script (available in the GitHub repo) for barcode correction. Barcodes were corrected in a similar manner as in the 10X Genomics Cell Ranger ATAC v. 1.0 software. In brief, barcodes were checked against the 10X Genomics whitelist. If a barcode was not on the whitelist, then we found all whitelisted barcodes within a hamming distance of two from the bad barcode. For each of these whitelisted barcodes, we calculated the probability that the bad barcode should be assigned to the whitelisted barcode using the phred scores of the mismatched base(s) and the prior probability of a read coming from the whitelisted barcode (based on the whitelisted barcode's abundance in the rest of the data). If there was at least a 97.5% chance that the bad barcode was derived from one specific whitelisted barcode, it was corrected to the whitelisted barcode.

Reads were mapped using bwa mem with flags '-I 200,200,5000 -M'. We used Picard MarkDuplicates to mark duplicates, and filtered to high-quality, non-duplicate autosomal read pairs using samtools view with flags '-f 3 -F 4 -F 8 -F 256 -F 1024 -F 2048 -q 30'. Quality control metrics were gathered on a per-nucleus basis using ataqv (v. 1.1.1) on the bam file with duplicates marked. In the case of the mixed rat and human snATAC-seq library, all reads were mapped to the hg19 and rn6 genomes separately, and then a nucleus was assigned as either rat or human by counting the number of high-quality, non-duplicate autosomal reads after mapping to either genome. If at least three times as many high-quality reads were present after mapping to one genome than to the other, the nucleus was assigned to either the rat or human sample as appropriate. In the case that fewer than three times as many high-quality reads mapped to one genome as to the other, the nucleus was not assigned to either species and was dropped.

For the two snATAC-seq libraries that contained a mix of nuclei from the two human individuals, we assigned nuclei to biological samples (and determined doublets) using demuxlet [114] with SNP calls from the bulk ATAC-seq libraries. To call SNPs on the bulk ATAC-seq BAM files, we first merged the two bulk technical replicate ATAC-seq BAM files for each individual, then filtered out reads with edit distance $>$ 2 from the hg19 reference. Used samtools mpileup (-R -Q 20 -d 10000 -E) on these two BAM files as input to bcftools call (-v -f GQ; v. 1.9). We then used bcftools filter to filter to those positions where both individuals had genotype quality (GQ) $>$ 90. This VCF file was used as input to demuxlet (option '–field PL'; git commit b7453fc, modified as described in GitHub issue #15).

When comparing aggregate snATAC-seq signal to bulk ATAC-seq signal (Fig. 4.1), we eliminated sequencing reads corresponding to nucleus barcodes that couldn't be matched to the 10X barcode whitelist, but otherwise processed it as bulk ATAC-seq data (i.e., marking duplicates ignoring cell-level information, and not filtering to quality nuclei).

To select quality nuclei from each library, we selected nuclei (barcodes) meeting the thresholds in Table 4.4.

4.5.7 Processing of snRNA-seq data

snRNA-seq data was processed using starSOLO (STAR v. 2.7.3a), which outputs the count matrices needed for most of the analyses [59]. To select quality nuclei from each library, we selected nuclei meeting the thresholds in Table 4.5. We used souporcell (as contained in the Singularity container downloaded from the souporcell GitHub on Dec. 10, 2019, and setting -k 2) to detect doublets in the libraries that were a mix of nuclei from two human individuals [98]. We additionally ran doubletfinder

library	min. reads	max. reads	min. TSS enrich.	max(max fraction reads from single autosome)
Human-rat mix (human nuclei)	25000	335319	3	0.15
Human-rat mix (rat nuclei)	35000	348820	3	0.15
no FANS, rep. 1	70000	342966	4.5	0.15
no FANS, rep. 2	50000	214062	4.5	0.15
20k nuclei	30000	159372	4.5	0.15
40k nuclei	15000	72704	4.5	0.15

Table 4.4: snATAC-seq nucleus QC thresholds

library	min.UMIs	max.UMIs	max. fraction mitochondrial
no FANS, rep. 1	2000	35000	0.004
FANS, rep. 1	1500	25000	0.009
no FANS, rep. 2	2000	35000	0.004
FANS, rep. 2	1500	25000	0.009
20k nuclei	1000	10000	0.007
40k nuclei	1000	10000	0.009

Table 4.5: snRNA-seq nucleus QC thresholds

(v. 2.0.2) [181] on each of the snRNA-seq libraries, and removed any nuclei that were called as a doublet by either souporcell or doubletfinder. When running Seurat (v. 3.0.2) for doubletfinder, we set selection.method = ‘vst’ and nfeatures = 2000, and used the top 20 PCs to find neighbors and resolution = 0.8 to find clusters [27, 269]. When calling the doubletFinder_v3 function, we selected the doubletfinder pK based on the maximum ’BCmetric’ after running the paramSweep_v3 function, set nExp assuming a 7.5% doublet rate (adjusting for the homotypic proportion as in the doubletfinder documentation example), and used the top 20 PCs.

4.5.8 Clustering with LIGER

Nuclei were clustered using LIGER (v. 0.4.2; with R v. 3.5.1 and Seurat v. 2.3.0) [300, 27, 269]. For snATAC-seq libraries, per-gene scores were computed by calculating the number of reads overlapping with each gene's promoter/gene body using bedtools intersect. Gene promoter/body were calculated based on NCBI annotation GTF files (NCBI Rattus norvegicus Annotation Release 106 and Homo sapiens Updated Annotation Release 105.20190906), filtered to include only protein-coding/lncRNA genes with source 'BestRefSeq'/BestRefSeq%2CGnomon'/'Curated Genomic'. Genes assigned to multiple chromosomes/strands were excluded, and then the regions for each gene were merged to get the gene body. Promoters were taken as the 3kb upstream of the TSS; after this, genes represented by multiple non-contiguous genomic stretches were excluded. For input to LIGER, all count matrices for a given modality and biological sample were concatenated together, so that there was 1 rat snATAC matrix, 2 human snATAC matrices, and 2 human snRNA matrices. For factorization, we used k = 15, lambda = 5, and nrep=5, using the smaller human snRNA matrix to select variable genes (as all the nuclei for that matrix were processed on a single day, and should therefore reflect less technical variation). For each of the downstream steps we dropped factors 3 and 5, as these had highly-loading ribosomal genes or showed relatively high specificity for one of the two omics modalities. For normalization, we set knnk (and small.clust.thresh) to 10 and resolution to 0.05, and centered the data. For the UMAP, we used n_neighbors = 15. We then called the clusterLouvainJaccard function to re-cluster cells using the normalized factors, with k = 17, and resolution = 0.05.

4.5.9 Per-cluster processing of snATAC-seq data

The filtered reads from all snATAC-seq nuclei in each cluster were merged using samtools merge. Peaks were called and bigwig files produced as described for the bulk ATAC-seq data. Peak files were filtered against blacklist files available from http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeMapability/wgEncodeDacMapabilityConsensusExcludable.bed.gz and http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeMapability/wgEncodeDukeMapabilityRegionsExcludable.bed.gz (hg19) (ENCODE Project Consortium, 2012) and https://github.com/shwetaramdas/maskfiles/tree/master/rataccessibleregionsmaskfiles/strains_intersect.bed for rn6 (Ramdas et al., 2019).

For analysis of rat peak overlap with human GWAS data, rat peaks were projected into the human genome using bnMapper (v. 0.8.6) and the chain file at http://hgdownload.cse.ucsc.edu/goldenpath/rn6/liftOver/rn6ToHg19.over.chain.gz.

4.5.10 Roadmap enhancer regression

We called peaks on the aggregate of the nuclei in each cluster, and then took the union of peaks across all clusters to generate a master peak list. We then used logistic regression to model, for each cluster and each Roadmap Epigenomics cell type in the Roadmap 15-state chromHMM model, the accessibility of each TSS-distal master peak ($>$ 5kb from a RefSeq TSS) in that cluster as a function of the posterior probability that that master peak is an enhancer in that Roadmap cell type according to the Roadmap chromHMM model [243]. Since the posteriors are given in 200 bp windows, and there are also 3 different enhancer states ('Genic enhancers', 'Enhancers', and 'Bivalent Enhancer'), multiple windows overlap with each master peak – the posterior for the master peak is therefore taken as the maximum of the 200 bp window posteriors, across all 3 of the enhancer states. The model coefficient was used as the (unnormalized) score for that Roadmap cell type in that cluster, and the normalized score was simply the score for that Roadmap cell type in that cluster divided by the max score across all cell types for that cluster.

For rat peaks, in addition to removing master peaks near TSS in rat coordinates, we additionally removed master peaks that were within 5 kb of a TSS after projecting into human coordinates.

4.5.11 Non-muscle cell type open chromatin annotations used in GWAS

To create the adipose open chromatin regions, we processed the three adipose ATAC-seq libraries from [29]. Adapter sequences were removed using Cutadapt (v. 1.12) [176] before mapping to hg19 with bwa mem (-I 200,200,5000 -M). Duplicates were marked using picard MarkDuplicates and BAM files were filtered using samtools view (-F 4 -F 256 -F 1024 -F 2048 -q 30) before converting to BED format (bamtools bamtobed) and calling peaks with MACS2 (–nomodel –shift -100 –seed 2018 –extsize 200 –broad –keep-dup all –SPMR). We then took the union of peaks across the three samples, keeping those merged peaks that appeared in at least two samples.

The beta cell ATAC-seq peaks were taken from [228]. We used the peaks called using all beta cell nuclei.

Common open chromatin regions were derived from the DNaseI hypersensitive sites from [184]. The DHS index from [184] was downloaded from `https://www.meuleman.org/DHS_Index_and_Vocabulary_hg38_WM20190703.txt.gz` on March 21, 2020. We lifted open chromatin regions from hg38 to hg19 using liftOver with the chain file from `http://hgdownload.cse.ucsc.edu/goldenpath/hg38/liftOver/hg38ToHg19.over.chain.gz` [102]. We then kept those that were labeled as 'tissue invariant' and that appeared in at least 500 of the 733 samples.

We also used open chromatin regions from [184] for adrenal gland, bone, brain, eye, gonad, gum, heart, kidney, large intestine, liver, lung, mammary, mesoderm, ovary, placenta, prostate, skin, small intestine, spinal cord, spleen, stomach, and umbilical cord. For each tissue, we took the non-cancerous samples labeled 'Primary' from that tissue and kept those DNaseI hypersensitive sites that appeared in at least 50% of the samples from that tissue.

4.5.12 UK Biobank GWAS enrichment

We downloaded UK Biobank GWAS summary statistics made available by the Benjamin Neale lab (v2 of their analysis, initially made public on August 1, 2018; `http://www.nealelab.is/uk-biobank/`) [270]. Specifically, we downloaded the 'both sex' GWAS summary statistic files listed in the 'UKBB GWAS Imputed v3 - File Manifest Release 20180731' spreadsheet available at `https://docs.google.com/spreadsheets/d/1kvPoupSzsSFBNSztMzl04xMoSC3Kcx3CrjVf4yBmESU/edit#gid=178908679` (downloaded on April 9, 2020). Because some traits may not be appropriate for such an enrichment analysis (because they are not strongly polygenic, because the phenotypes are untrustworthy, etc.), we kept only traits deemed as 'high confidence' and with estimated heritability > 0.01 (and z-score > 7) based on the Neale Lab's own LD score regression heritability analysis of the GWAS results. Their rating criteria are described on their UKBB LDSC GitHub page (`https://nealelab.github.io/UKBB_ldsc/confidence.html`) and their LD score regression results (with confidence ratings) were downloaded from `https://www.dropbox.com/s/ipeqyhrpdqav5uh/ukb31063_h2_all.02Oct2019.tsv.gz?dl=1`. For each trait, we used the 'primary' GWAS result, as indicated in that file. Any traits that did not have a combined male and female GWAS analysis were dropped. The creatinine GWAS highlighted in the text was trait 30700_irnt ('Creatinine (quantile)').

The LDSC software package (v. 1.0.1) includes a 'baseline' model with 59 categories derived from 28 genomic annotations [71, 80]. Many of these annotations are cell type agnostic; e.g. a SNP's minor allele frequency does not change between cell types. However, other annotations in the baseline model are not cell type agnostic; for example, the FANTOM5 enhancer annotation is derived from experiments performed on a range of different cell types, and may change substantially if the cell types used to create the annotation were to change. When performing the UK Biobank GWAS enrichments, we utilized the cell-type agnostic annotations from the

LDCS baseline model. In order to reduce the likelihood of model misspecification, we then added common open chromatin regions and open chromatin regions from a range of cell types. Specifically, we added (1) beta cell ATAC-seq peaks, (2) adipose ATAC-seq peaks, (3) DNase-seq peaks derived from the 22 tissues/organs listed above, and (4) the ATAC-seq peaks from all seven of our snATAC-seq cell types. The various annotation files (regression weights, frequencies, etc.) required for running LDSC were downloaded from `https://data.broadinstitute.org/alkesgroup/LDSCORE`. LD scores were calculated using the Phase 3 1000 Genomes data, keeping only the HapMap3 SNPs as recommended by the LDSC authors and using only SNPs with minimum MAF of 0.01. GWAS summary statistics were prepared for LDSC using the munge_sumstats.py script, with option –merge-alleles w_hm3.snplist (where w_hm3.snplist is the file in the data download). When running the regression, we required a minimum MAF of 0.05, and utilized the Phase 3 1000 Genomes SNP frequencies/weights.

4.5.13 T2D and fasting insulin GWAS enrichment

We used the T2D (BMI unadjusted) and fasting insulin (BMI adjusted) GWAS summary statistics from [172] and [175], respectively.

Because the cell types relevant to T2D are generally thought to be pancreatic beta cells, adipose, muscle, and liver, we performed enrichments using each of these cell types, common open chromatin, and the cell type-agnostic LDSC baseline annotations. First, for each of these muscle/beta cell/adipose/liver cell types, we ran one model containing the open chromatin from that cell type, the common open chromatin regions, and the cell type-agnostic LDSC baseline annotations. Then, we ran one joint model containing all of those cell types and annotations. LDSC parameters were the same as for the UK Biobank GWAS enrichments.

4.5.14 T2D GWAS locus genome browser screenshots and peak overlaps

All signal tracks in the genome browser were created by converting the normalized bedgraph files output by MACS2 to bigwig files using bedGraphToBigWig (v. 4).

Processing and provenance of adipose ATAC-seq and beta cell ATAC-seq is described above. The 10 bulk islet libraries were from [228]. These libraries were processed as described in that manuscript, except we did used the 10% FDR peak set from peak calling on the unsubsampled libraries.

4.5.15 Predicting SNP regulatory impact

We used the lsgkm package modified by the Kundaje lab with gkmexplain (`https://github.com/kundajelab/lsgkm`; commit c3758d5bee7) [82, 150, 265]. For each cell type, we took the 150 bps on either side of the summits of the top 40,000 narrowPeaks (by p-value) as the positive sequences for gkmSVM. To generate negative sequences, we took windows across the genome (step size = 200), removed those containing Ns, overlapping hg19 blacklists, overlapping any FDR 10% broadPeaks from that cell type, or having repeat content $> 60\%$, and then for each positive sequence selected a negative sequence with matching GC content and repeat content (repeat content was calculated based on the hg19 simpleRepeat table from the UCSC genome browser [122, 31], downloaded on March 29, 2020, which contains simple tandem repeats annotated by Tandem Repeats Finder [15]; GC content and repeat content for the negative sequence was required to be within 2% of that of the positive sequence; in the case that no such negative sequence could be found, the positive sequence was dropped from the analysis). We held out 15% of sequences as test data, and trained the gkmSVM model on the remaining 85% of sequences, setting $l = 10$ and $k = 6$ and using the gkm kernel. Using this model and deltaSVM [151], we predicted the effect of all autosomal 1000 Genomes phase 3 SNPs (downloaded on May 27, 2015 from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502) (1000 Genomes Project

Consortium et al., 2015). For each muscle cell type, deltaSVM scores were converted to z-scores based on the distribution of scores across all SNPs for that cell type. We additionally passed the gkmSVM model to gkmexplain to generate importance scores for sequences containing the ref/alt alleles.

4.5.16 Overlap of SNPs and peaks with ENCODE candidate cis-regulatory elements

The set of 1,310,152 candidate cis-regulatory elements in ENCODE's 'Registry of candidate Regulatory Elements' (in hg19 coordinates) were fetched from the ENCODE web portal on April 7, 2020 [64, 65].

4.5.17 Locuszoom plots

Locuszoom plots were created for the DIAMANTE T2D GWAS summary statistics with the locuszoom standalone v. 1.4, using the Nov. 2014 EUR 1000 Genomes data included in the download (–pop EUR –source 1000G_Nov2014) [221].

4.5.18 PICS

We used the online PICS tool [70] (`https://pubs.broadinstitute.org/pubs/finemapping/pics.php`) with the EUR LD structure. The tool was accessed on April 13, 2020.

4.5.19 Motif scan

The MEF2 motif scan was performed using FIMO (v. 5.0.4) [90] with a background model calculated from the hg19 reference genome.

4.6 Acknowledgements

We wish to thank the University of Michigan Advanced Genomics Core for their assistance in generating the snRNA-seq and snATAC-seq libraries, and the University of Michigan Flow Cytometry Core for their help performing FANS. We are grateful to the Benjamin Neale lab for providing their UK Biobank GWAS and LDSC results to the scientific community. We also thank members of the Parker lab for their helpful feedback.

4.7 My contributions

Nandini Manickam generated the bulk ATAC-seq data and performed the nuclear isolations for the single-nucleus datasets. Arushi Varshney and Vivek Rai processed existing islet and adipose ATAC-seq data. Jeremy Kaplan was a Master's student in the lab and did much of the work involved in setting up gkm-SVM models. Claudia Lalancette from the University of Michigan Epigenomics Core helped coordinate production of the single-nucleus data. Charles Burant provided the rat muscle sample, and Katherine Gallagher provided human muscle samples. I thank each of these individuals for their contributions to this project, as well as Steve for his ever-insightful advice and guidance. I performed all computational processing and analyses of the data not attributed above to others, and contributed the majority of the text. While working on this project, I was funded by a University of Michigan Rackham Predoctoral Fellowship as well as the University of Michigan Genome Sciences Training Program (grant T32 HG00040 from the National Human Genome Research Institute of the N.I.H.)

CHAPTER V

Association of Skeletal Muscle Molecular Traits with Genetic Variants and Running Capacity in a Rat Model for Aerobic Exercise Capacity

5.1 Abstract

In humans, high aerobic exercise capacity is associated with decreased susceptibility to many diseases and decreased mortality overall. Determining the genes and gene regulatory elements related to exercise capacity may clarify the molecular relationships underlying this association. Here, we utilize a rat model of aerobic exercise capacity to search for genes and non-coding regulatory elements linked to exercise capacity, as well as identify thousands of genetic variants associating with chromatin accessibility.

5.2 Introduction

Complex, polygenic diseases account for a large fraction of deaths and healthcare spending in developed nations [303, 220, 81], and are a growing burden in developing nations [220, 81]. Aerobic exercise capacity (AEC) is one strong predictor of mortality [147, 174, 106, 304, 297] and risk of common diseases like cardiovascular disease

[147, 106, 304, 297] and cancer [106]. AEC is a highly heritable trait [19, 258] and understanding the genetic variants and molecular traits associated with AEC could clarify the molecular pathways underlying the observed relationship between AEC and many common complex diseases.

One recently-developed rat model for AEC provides further compelling evidence for the link between AEC and other health phenotypes and is a promising system for investigating the molecular traits associated with AEC [130, 131, 132]. This rat model is comprised of two rat lines that are divergent in their running capacity, referred to as high capacity runners (HCRs) and low capacity runners (LCRs). Generated via artificial selection for intrinsic (untrained) running capacity as measured by treadmill tests, the HCRs and LCRs show a striking ($\sim$ 7-fold [132]) difference in running capacity. Although artificial selection is performed based on running distance, the HCRs and LCRs have developed remarkable differences in many other traits as well, including body mass [130] (lower in HCRs relative to LCRs), susceptibility to insulin resistance [190] (lower in HCRs relative to LCRs) and lifespan (higher in HCRs relative to LCRs) [133]. F2 rats derived from HCRs and LCRs (created by mating the HCRs with the LCRs, and then mating offspring pairs) show a wide distribution of running capacity [239] and represent an interesting cohort for studying associations between running capacity and molecular traits.

Here, we perform ATAC-seq and RNA-seq on up to 150 F2 rats and examine associations between chromatin accessibility, gene expression, and running capacity. We additionally associate chromatin accessibility with genetic variants to find thousands of chromatin quantitative trait loci (caQTL).

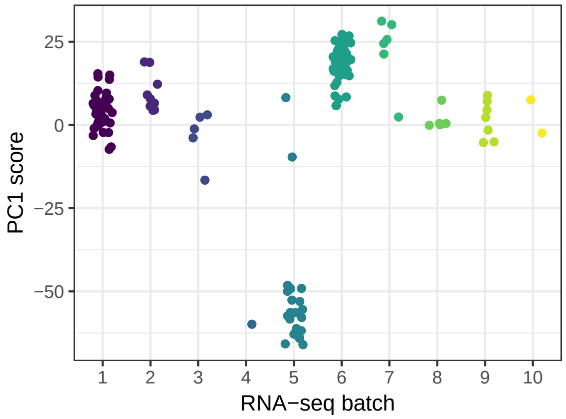


Figure 5.1 RNA-seq PC1 separates one batch from the others.

5.3 Results

5.3.1 Transcriptomic and epigenomic profiling of F2 rat skeletal muscle samples

We performed bulk RNA-seq on 150 F2 rats and bulk ATAC-seq on 141 of the 150 rats. After removing samples swaps and performing quality control (see section 5.5), 143 RNA-seq and 129 ATAC-seq libraries remained. Principal component analysis (PCA) on the log2(FPKM) gene expression matrix showed signs of a batch effect (PC1; Fig. 5.1), while other top PCs correlated with sex. Top PCs from a PCA on the matrix of chromatin accessibility (log2(RPKM) values for each autosomal ATAC-seq peak) correlated with several ATAC-seq quality control metrics, consistent with variance introduced by technical biases (Fig. 5.2).

5.3.2 Running capacity association with gene expression

To find genes that may play a role in running capacity, we modeled the relationship between running capacity (best running distance) and the expression of each autosomal gene. Because some of the rats are siblings and siblings are expected to show increased similarity in both running capacity as well as gene expression relative to non-siblings, we expected p-values from linear models may be inflated; therefore, we

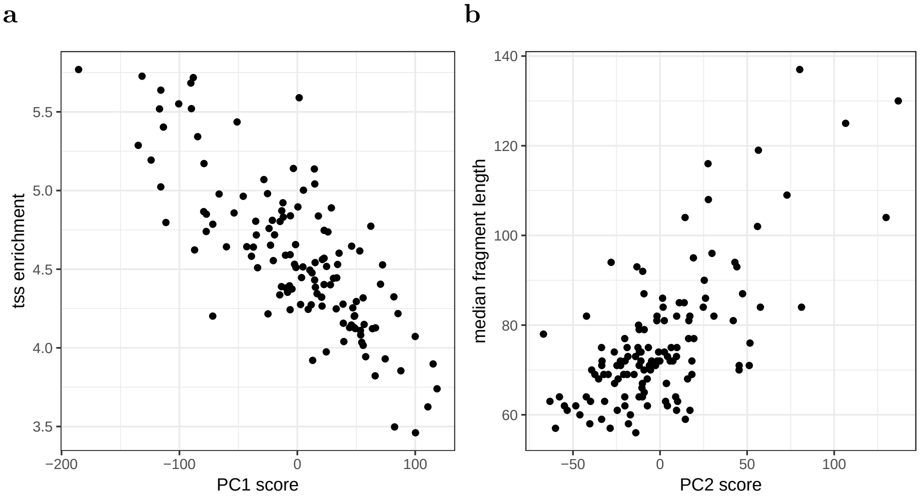


Figure 5.2 Correlation between top ATAC-seq PCs and quality control metrics. (a) PC1 correlates with TSS enrichment, a measure of signal-to-noise. (b) PC2 correlates with median fragment length.

ran several linear models (not adjusting for relatedness in any way) and linear mixed models (LMMs; using a genetic relatedness matrix estimated from all genotyped SNPs as the correlation matrix in a variance component to adjust for relatedness; see methods in section 5.5.8) and compared the results (Fig. 5.3). Because female F2 rats tend to have greater running capacity than male F2 rats (Fig. 5.4), we included sex as a fixed effect in all models and examined the impact of adding other possible covariates (body weight, sequencing run, and several RNA-seq QC metrics including the fraction of reads assigned to the top 100 expressed genes (summarizing the distributions shown in Fig. 5.12) and a measure of 3' bias) (Fig. 5.3). Most linear models and linear mixed models show substantial inflation; however, we found that the linear mixed model including the fraction of reads assigned to the top 100 expressed genes as a covariate appeared less inflated than all other models, including the corresponding linear model (Fig. 5.3). We therefore proceeded to examine linear mixed models with sex, fraction of reads assigned to the top 100 expressed genes, and either sequencing run or body weight as an additional covariate (Fig. 5.5). Adding body weight as a covariate appeared to slightly reduce inflation over the first half of the Q-Q plot (though the Q-Q plots are broadly similar; Fig. 5.5); therefore, we chose to proceed with the linear mixed model containing gene expression, sex, fraction of reads assigned to the top 100 expressed genes, and body weight as fixed effects (in addition to the random effect controlling for genetic relatedness). The tail of the Q-Q plot may show hints of some real signal; however, none of the genes are significantly associated with running capacity (at FDR $< 5\%$). This result is perhaps unsurprising, as a previous analysis of running capacity association with gene expression in a larger cohort of 409 F2 rats [230] identified only 14 associated genes (with considerable inflation in the associated Q-Q plot).

The distribution of running capacity heritability (h^2) point estimates based on the proportion of variance explained by the random effect in the LMM model (one model per gene) is displayed in Figure 5.6. The median estimate was 0.55, similar to the running capacity heritability estimate from GCTA using only best running distance and the genetic relatedness matrix for the 143 rats with pass-QC RNA-seq data (0.49 ± 0.15).

Despite the absence of significant genes after multiple testing correction, we used RNA-Enrich [148] to determine if genes with more extreme p-values (from the LMM) are enriched for certain biological processes. RNA-Enrich is particularly suited for this analysis, as it does not impose a p-value threshold in the analysis; rather, it computes a test statistic for each gene ontology (GO) term using -log10(p) for all of the tested genes included in that GO term. We found that genes with more extreme p-values for being positively associated with running capacity were strongly enriched for fatty acid oxidation processes (Fig. 5.7). This result is consistent with several previous analyses on the HCR/LCR rats, which found that HCRs have an increased capacity for fatty acid oxidation [204], genes expressed more highly in HCRs than LCRs are enriched for fatty acid oxidation pathways [238], and genes more strongly associated with running capacity in the F2s (in extensor digitorum longus (EDL)

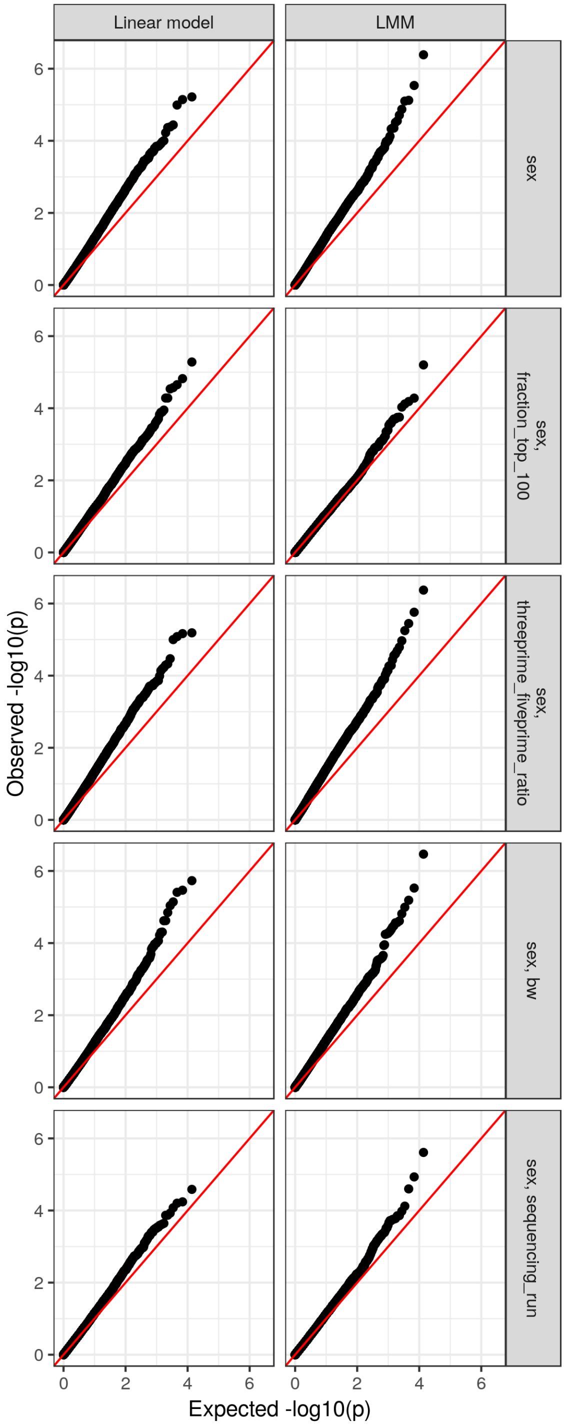


Figure 5.3 Q-Q plots for gene expression association with running distance. We ran linear models (no adjustment for relatedness between rats) or linear mixed models (‘LMM’; using a variance component to adjusted for relatedness using the genetic relatedness matrix). Facet labels on the right indicate fixed effects included in the model (in addition to expression of the gene); bw = body weight, fraction_top_100 = fraction of reads assigned to the top 100 expressed genes. Most linear models and linear mixed models show substantial inflation; however, linear mixed models with the fraction of reads assigned to the top 100 expressed genes showed reduced inflation relative to other models.

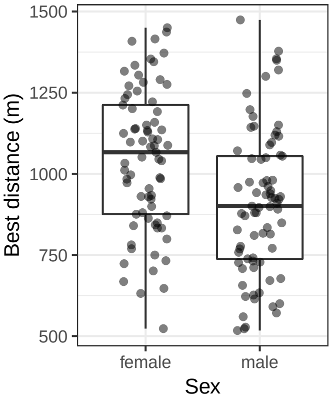


Figure 5.4 Best running distance by sex. Females show, on average, greater running capacity.

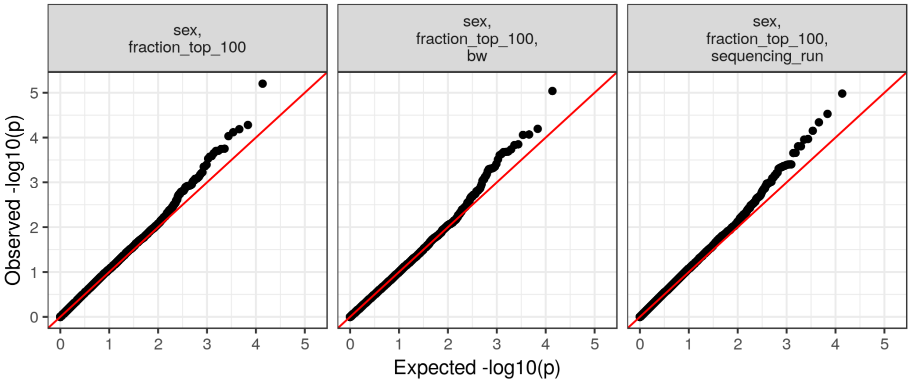


Figure 5.5 Q-Q plots for gene expression association with running distance (sex + fraction_top_100 + additional covariate. Facet labels indicate fixed effects included in the model (in addition to expression of the gene); bw = body weight, fraction_top_100 = fraction of reads assigned to the top 100 expressed genes.

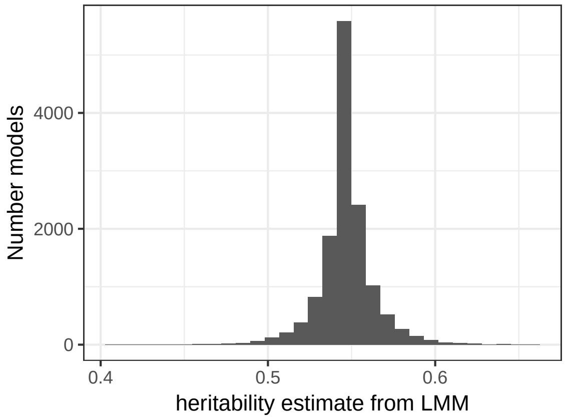


Figure 5.6 Distribution of running capacity heritability estimates from linear mixed model (LMM) for gene expression association with running distance.

muscle) are enriched in fatty acid oxidation pathways [230].

One possibility is that the effect sizes of genes associated with running capacity are too small to be detected in our analysis. If this were the case, we might at least expect to see that genes with a positive coefficient in our model (i.e., tilting towards positive association with running capacity, even though the association is not significant) also tilt (again, not necessarily significantly) towards higher expression in HCRs than in LCRs. To see if such an effect is present, we downloaded previously published HCR vs. LCR EDL muscle gene expression log fold changes [238] and compared the sign of the log fold change for each gene in that data set to the sign of each gene's (non-significant) coefficient for association with running capacity in our model. To enrich for signal, we restricted this analysis to genes that were nominally significant ($p < 0.05$) in at least one of the analyses. Though the relationship is unsurprisingly noisy, we found that genes with a positive HCR/LCR log fold change were more likely to have a positive coefficient in our model, and genes with a negative HCR/LCR log fold change were more likely to have a negative coefficient (Fig. 5.8). This is consistent with the hypothesis that there are genes significantly associated with running capacity that we lack the power to detect in our analysis.

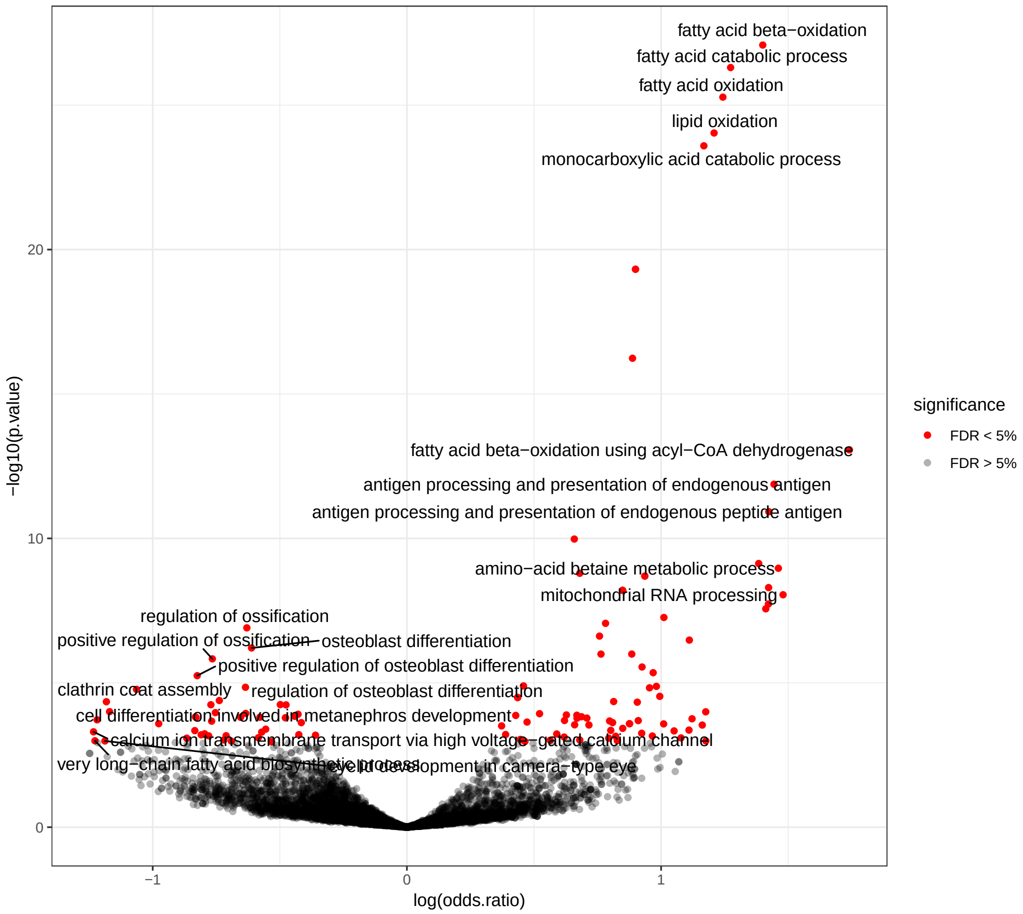


Figure 5.7 Gene ontology enrichment analysis using RNA-Enrich.

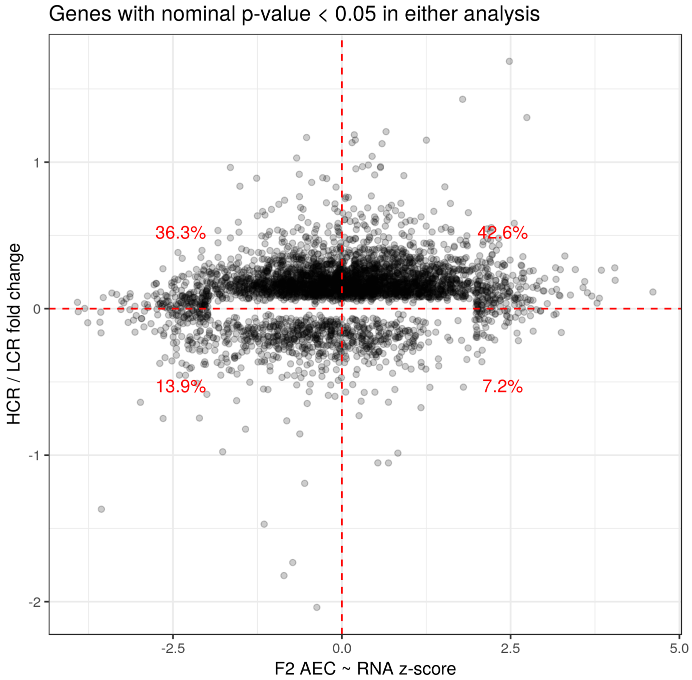


Figure 5.8 HCR/LCR log fold change vs gene expression coefficient in the running capacity ∼ gene expression analysis using F2 RNA-seq data. Percentages represent the percentage of genes in each quadrant. Only genes with p < 0.05 in at least one of the analyses are displayed.

5.3.3 Running capacity association with chromatin accessibility

The relative activity of enhancers or other gene regulatory elements could mediate changes in running capacity. While we do not have a direct measure of enhancer activity in these samples, chromatin accessibility is often used as an (imperfect) proxy for regulatory element activity. Furthermore, chromatin accessibility differences between rats could signal priming of regulatory elements to become active upon a change in exercise state, enabling some rats to respond to the initiation of exercise more efficiently than others. We therefore used linear models (not controlling for relatedness between rats) or LMMs (controlling for genetic relatedness between rats; see methods in section 5.5.9) to test for a relationship between chromatin accessibility and best running distance for 200,305 ATAC-seq peaks. We included sex as a fixed effect in all models and examined the impact of adding other possible covariates (body weight, sequencing run, and several ATAC-seq QC metrics including TSS enrichment (quantifying signal-to-noise in the library) and median fragment length; Fig. 5.9). Linear models do not appear more inflated than linear mixed models, and regardless of the type of model (linear or linear mixed model) and the covariates included, there is little evidence of chromatin accessibility association with running capacity; after FDR correction, no ATAC-seq peaks were significantly associated (FDR $< 5\%$) with running capacity in any of the models.

5.3.4 Variants associating with chromatin accessibility

Non-coding variants may act on a phenotype by perturbing regulatory element function. Although previous GWAS on running capacity in the F2 rats did not identify any significantly associated variants [230], identifying variants that correlate with changes in chromatin accessibility may assist future genetic analysis of the HCR/LCR model system. Therefore, we used RASQUAL [138] to detect caQTLs, controlling for sex, weight, sequencing run, and ATAC-seq PCs 1-3 (Fig. 5.10). We tested all 684,421

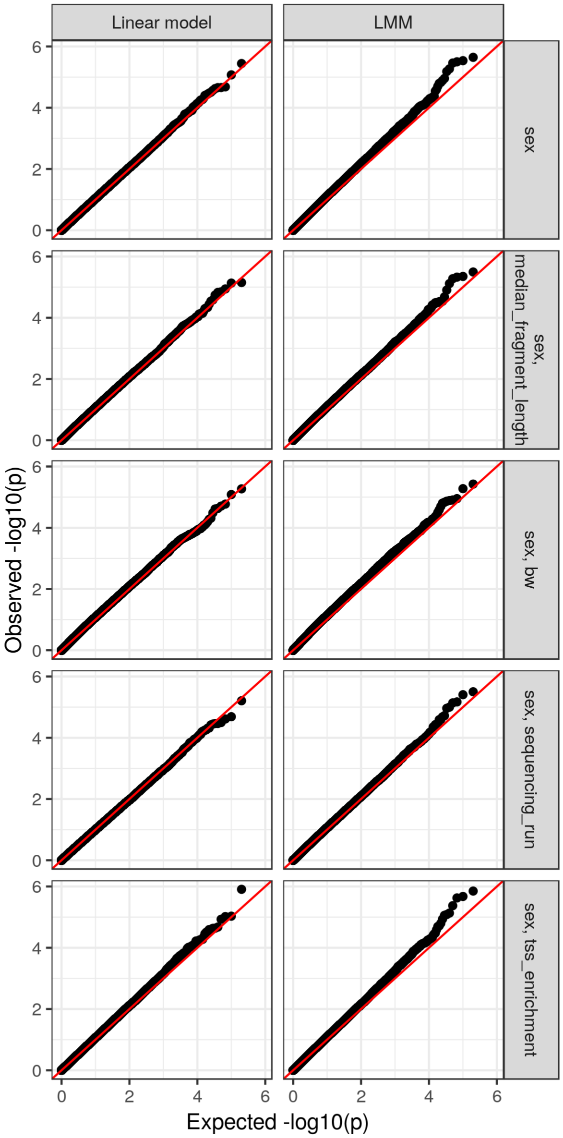


Figure 5.9 Q-Q plots for chromatin accessibility association with running distance. We ran linear models (no adjustment for relatedness between rats) or linear mixed models ('LMM'; using a variance component to adjusted for relatedness using the genetic relatedness matrix). Facet labels on the right indicate fixed effects included in the model (in addition to the accessibility of the peak); bw = body weight. Inflation is minimal regardless of the type of model and covariates used, and for all models no ATAC-seq peaks show significant association (FDR <5%) with running capacity.

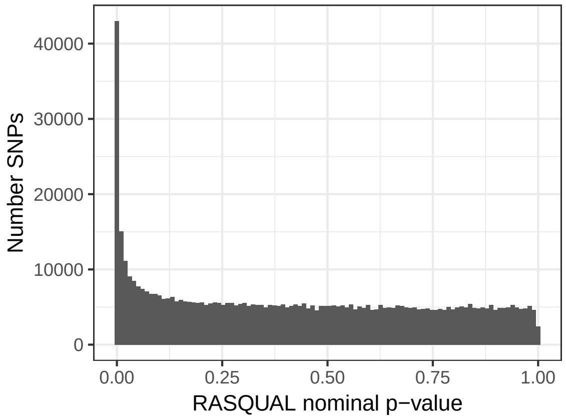


Figure 5.10 Distribution of caQTL nominal p-values.

SNP-peak pairs for which the SNP was within 10kb of the peak; this included 110,371 unique ATAC-seq peaks. LD between neighboring SNPs results in correlated p-values when they are tested against the same peak; therefore, to correct for multiple tests, we used eigenMT [51] to estimate the number of independent tests performed for each ATAC-seq peak and perform Bonferroni correction of p-values for each peak, and then performed Benjamini-Hochberg correction [13] across all peaks (using the top SNP p-value for each peak) to determine which peaks were significantly associated (FDR $<$ 5%) with at least one SNP. This procedure resulted in 4,477 peaks that were significantly associated with at least one SNP. An example locus, where SNP chr4:80860957 is a caQTL for the peak it is in ($p = 1.6 \times 10^{-26}$), is shown in Fig. 5.11; the A allele of this SNP correlates with increased chromatin accessibility. This peak sits 10.4kb from the nearest gene (a lncRNA, AABR07060560.3).

5.4 Discussion

Here, we first examined the relationship between running capacity and gene expression to find genes that may impact running capacity. While no genes were statistically significant at 5% FDR, we found genes with (non-significant) positive coefficients for association with F2 running capacity tended to lean towards higher expression in

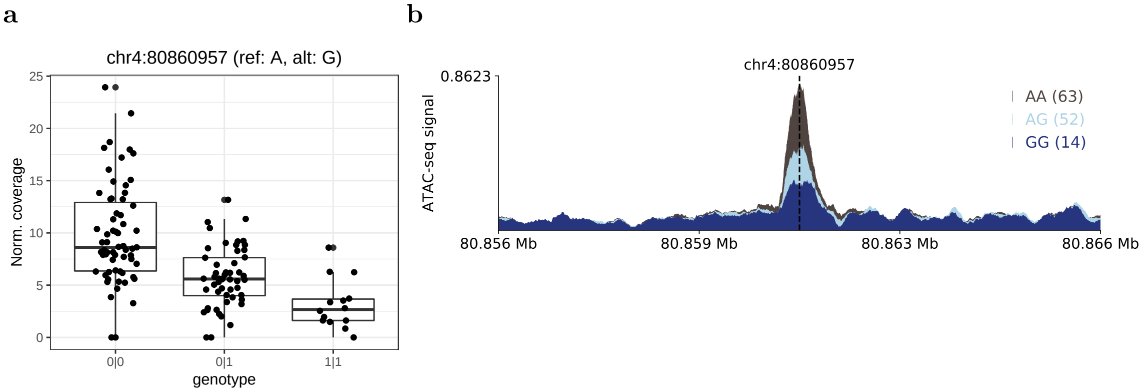


Figure 5.11 Example caQTL SNP-peak pair. (a) Read counts at SNP chr4:80860956 (caQTL p-value = 1.6×10^{-26}), normalized by the total number of reads in peaks for each sample. (b) Average ATAC-seq signal in the region around SNP chr4:80860956, stratified by rat genotype at the SNP. ATAC-seq signal is scaled to reads per million. The gene nearest to this locus is a lncRNA, AABR07060560.3, 10kb away.

HCRs than in LCRs and the reverse were true for genes with negative coefficients, suggesting there may be a substantial number of associated genes that we simply lack the power to detect using traditional FDR thresholds at the current sample sizes. Regardless of the lack of significant genes, a GO analysis suggested that genes with more extreme p-values are enriched for fatty acid oxidation, consistent with previous findings [204, 238, 230]. Next, we investigated the relationship between running capacity and chromatin accessibility. We found little evidence of association in this analysis – for all models tested, the distribution of p-values more-or-less follows the null expectation. This is evidence against the hypothesis that priming of gene regulatory elements in the resting state may play a factor in determining running capacity, although given the high multiple testing burden ($\sim$ 200k ATAC-seq peaks) and small sample size here (129 samples), it is of course possible that we are hampered by low power to detect small effects. There are additional tweaks that could be made to the models presented here that may improve results; for example, including an additional variance component in the linear mixed models to control for covariance in transcriptomes or chromatin accessibility across rats.

The above results, particularly for gene expression, are consistent with previous analyses suggesting that running capacity is a highly polygenic trait. The most recent QTL analysis on running capacity in these F2 rats (using $\sim$ 615 rats) did not find any significantly associated variants, and an association analysis of gene expression with running capacity in the same work (n = 409 F2 EDL skeletal muscle samples) identified relatively few significantly associated genes (14 with FDR $<$ 5%) [230]. With our limited sample sizes (143 for RNA-seq and 129 for ATAC-seq), it will be difficult to detect many small effects. This is a common challenge in the study of polygenic traits; methods that aggregate information across genomic loci or biological pathways might help overcome this, as demonstrated by the GO enrichment analysis.

All of these analyses were performed using bulk RNA-seq and bulk ATAC-seq with samples from rats at rest. It is possible that additional and/or stronger molecular trait associations with running capacity would be apparent under different physiological conditions (e.g., while the rats are running), or that some associations are cell type specific and may be difficult to detect without, for example, single-nucleus data. RNA-seq and ATAC-seq data is available for 128 HCR/LCR rats under a variety of exercise conditions; the data from that cohort will clarify the extent to which gene expression and chromatin accessibility change with exercise, and perhaps those results will help direct more targeted analyses in the F2 rats. We also plan to generate snRNA-seq and snATAC-seq data using these F2 samples in the future.

Lastly, we performed a caQTL analysis and identified 4,477 ATAC-seq peaks associated with at least one SNP. In isolation the caQTL results are not informative about the trait of interest, but these results may be helpful in interpreting future genetic findings in this model system or in other rat models.

5.5 Methods

5.5.1 Treadmill test

To determine the aerobic exercise capacity of each rat, a shock grid-assisted treadmill test is performed [130]. The treadmill speed at the beginning of the test is 10 meters/minute, and the speed is increased 1 meter/minute every two minutes. The treadmill slope is 15 degrees for the duration of the test. A shock grid sits at the bottom of the treadmill. Once the rat ceases to run and stays on the shock grid for two seconds, the rat is said to be exhausted.

5.5.2 RNA-seq library generation

RNA was isolated from 150 F2 gastrocnemius muscle samples, and stranded RNA-seq libraries were generated using poly(A) selection. Libraries were sequenced using paired-end, 150 bp reads.

5.5.3 ATAC-seq library generation

We pulverized 149 F2 gastrocnemius muscle samples. These samples were divided into 47 batches comprising of 6 samples. Each of the samples weighing around 50 mg were weighed out from the respective pulverized tissue samples. Each of the six samples was resuspended in 1 mL of ice-cold PBS, to perform six different nuclei isolations. Nuclei were isolated using a modified version of the ENCODE protocol [65]. The nuclei were counted in a Countess II FL Automated Cell Counter, and the appropriate volume of the suspension for 50K nuclei was spun down and used for the downstream transposition reaction. Libraries were sequenced using paired-end, 51 bp reads.

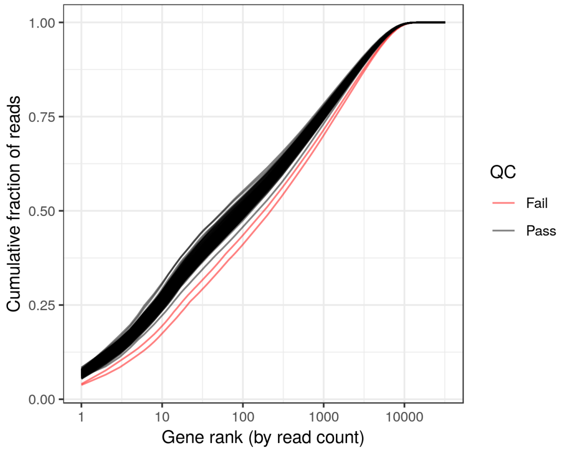


Figure 5.12 Cumulative read gene assignment diversity. Genes were ranked according to read count (high to low; X-axis), and the cumulative fraction of reads assigned to the top X genes was calculated (Y axis). Two outlier libraries, highlighted in red, were dropped from downstream analysis.

5.5.4 RNA-seq processing and quality control

Reads were mapped to the rn6 using STAR (v. 2.5.4b; options –outSAMunmapped Within KeepPairs) ([59]). QoRTs (v. 1.0.7) ([96]) was run on the resulting BAM files to generate QC metrics as well as gene fragment counts for downstream analysis (using all uniquely mapped, properly-paired read pairs). We identified two libraries that were outliers across several QoRTs QC metrics, including in cumulative gene expression (Fig. 5.12); these libraries were excluded from downstream analysis.

5.5.5 ATAC-seq processing and quality control

Adapters were trimmed using cta (v. 0.1.2; `https://github.com/ParkerLab/cta`). Reads were mapped to rn6 using bwa mem (-I 200,200,5000 -M; v. 0.7.15-r1140) [156]. Duplicates were marked using picard MarkDuplicates (v. 2.21.3; `https://broadinstitute.github.io/picard/`). We used samtools to filter to high-quality, properly-paired autosomal read pairs (-f 3 -F 4 -F 8 -F 256 -F 1024 -F 2048 -q 30; v. 1.9 using htslib v. 1.9) [157]. To call peaks, we used bedtools bamtobed [225] to convert to a bed file (v. 2.27.1) and then used that file as input to MACS2 callpeak

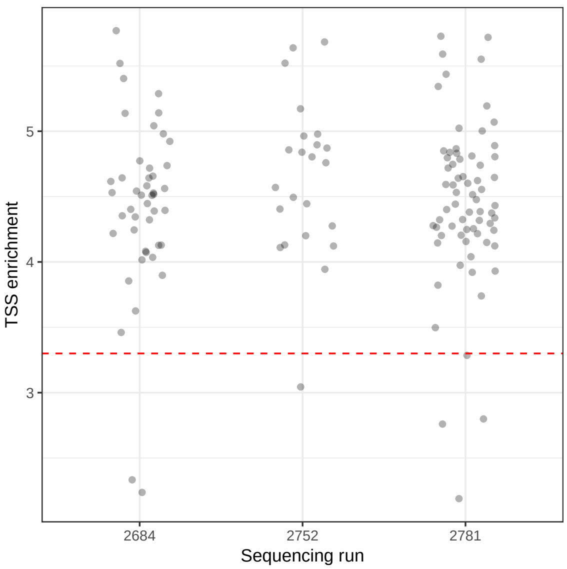


Figure 5.13 TSS enrichment for F2 ATAC-seq libraries. Libraries with TSS enrichment < 3.3 (denoted by the dashed red line) were dropped from downstream analysis.

(–nomodel –shift -100 –seed 762873 –extsize 200 –broad –keep-dup all –SPMR; v. 2.1.1.20160309) [316]. Peaks were filtered against the rn6 mask file from [231] (downloaded from `https://github.com/shwetaramdas/maskfiles/tree/master/rataccessibleregionsmaskfiles/strains_intersect.bed`). To visualize the signal, we converted the bedgraph files output by MACS2 to bigwig files using bedGraphToBigWig (v. 4) [123].

Basic ATAC-seq quality control was performed using ataqv [202]. In addition to the five sample swaps, we dropped seven libraries showing low TSS enrichment (Fig. 5.13), leaving 129 libraries for downstream analysis.

To generate a set of peaks for downstream analysis, we randomly sampled 5M reads from each rat and called peaks as described above on the resulting BAM file. We then kept only those peaks supported by a peak call in at least 5 of the individual samples; this resulting in 205,315 peaks.

5.5.6 Sample swaps

We checked for sample swaps in the ATAC-seq and RNA-seq data using the directly genotyped variants and MBV from the QTLtools software package [72, 53]. The same five swaps were present in the ATAC-seq and RNA-seq data; these libraries were removed for downstream analysis.

5.5.7 Genotyping

Genotyping was performed using the custom array from [237]. SNPs were lifted to the rn6 genome assembly, after which 378,951 SNPs remained.

5.5.8 Running capacity association with gene expression

To model the relationship between gene expression and best running distance accounting for genetic relatedness we used linear models (R's lm function):

$$Y = X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma_e^2 I)$$

and linear mixed models (as implemented in the lme4qtl R package [321], fitting the model using restricted maximum likelihood (REML)):

$$Y = X\beta + u + \epsilon, \quad u \sim N(0, \sigma_g^2 G), \epsilon \sim N(0, \sigma_e^2 I)$$

where Y is running distance, X is an incidence matrix, β is a vector of fixed effects, and G is a kinship matrix generated using EMMAX (Balding-Nichols matrix) [113] with the full set of autosomal SNPs, normalized as described in equation 5 of [113]. Gene expression values were converted to counts per million (CPM) and then inverse normalized. As discussed in section 5.3.2, for downstream analyses we used a linear mixed model with gene expression, sex, body weight, and the fraction of reads assigned to the top 100 expressed genes (summarizing the distributions shown in Fig.

5.12) as fixed effects, and a random effect to account for genetic relatedness (based on the kinship matrix). We tested all autosomal genes with at least 2 counts in 75% of the libraries.

5.5.9 Running capacity association with chromatin accessibility

To model the relationship between chromatin accessibility and best running distance accounting for genetic relatedness we used linear models (R's lm function):

$$Y = X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma_e^2 I)$$

and linear mixed models (as implemented in the lme4qtl R package [321], fitting the model using restricted maximum likelihood (REML)):

$$Y = X\beta + u + \epsilon, \quad u \sim N(0, \sigma_g^2 G), \epsilon \sim N(0, \sigma_e^2 I)$$

where Y is running distance, X is an incidence matrix, β is a vector of fixed effects, and G is a kinship matrix generated using EMMAX (Balding-Nichols matrix) [113] with the full set of autosomal SNPs, normalized as described in equation 5 of [113]. Peak read counts were converted to counts per million (CPM) and then were inverse normalized for each ATAC-seq peak. We tested all 200,305 peaks with more than 5 counts in at least 75% of the libraries.

5.5.10 Heritability estimation using GCTA

For estimating h^2 using GCTA, we first used GCTA (v. 1.93.2 beta) to generated a genetic relatedness matrix using all genotyped autosomal SNPs (i.e., no thinning of genotypes) with MAF > 1%, and then applied GCTA's GREML function using best distance as the phenotype and including no additional covariates.

5.5.11 caQTL analysis

We used RASQUAL (GitHub commit 1cdd0a0) [138] for the caQTL analysis. We included sex, weight, sequencing run, and ATAC-seq PCs 1-3 as covariates, and tested all SNPs within 10kb of each peak. For within-peak correction of p-values, we used eigenMT (downloaded from `http://montgomerylab.stanford.edu/resources/eigenMT/_downloads/eigenMTwithTestData.tgz` on May 15, 2020).

5.6 Acknowledgements and author contributions

Nandini Manickam generated the ATAC-seq libraries and Mary Treutelaar isolated RNA for the RNA-seq libraries. This project benefited from the input of each of my thesis committee members. I thank Charles Burant for supplying the muscle samples, and thank Charles Burant as well as Jun Li for informative discussions about the HCR-LCR rat model system and their interesting physiological traits and genetics. I thank Hyun Min Kang for advice about controlling for relatedness when modeling molecular trait - phenotype associations, and Jonathan Terhorst for several helpful discussions about the results of each analysis and strategies for analyzing polygenic traits more broadly. And as in each of my other dissertation chapters, Steve was generous with his time and brainpower in helping me move this project forward. I performed all of the bioinformatic analyses discussed here and wrote the text, with input from Nandini (who supplied some of the experimental methods) and Steve. During my work on this project, I was funded by a Rackham Predoctoral Fellowship from the University of Michigan Rackham Graduate School.

CHAPTER VI

Conclusion

In this thesis I have integrated epigenomic and transcriptomic data with genetic data to examine monogenic (chapter II) and polygenic (chapters IV, V) traits in human, mouse, and rat. I have also explored the diversity of public ATAC-seq datasets and the variability in ATAC-seq results introduced by several technical variables (chapter III). Each of these analyses generated new information or resources, but clearly none of them close the door on the problems they are attempting to tackle or the biological traits and systems they are investigating. Together they display the value of epigenomic and transcriptomic data in understanding monogenic and polygenic traits, and to varying degrees they also highlight new hypotheses and research directions, and current challenges.

In the Danforth (*Sd*) mouse project presented in Chapter II, I investigated the genome-wide changes in the transcriptome and epigenome of E9.5 *Sd* mouse tailbuds. The results are a demonstration of both the advantages of multi-omic data integration as well as the limits and challenges of integrating ATAC-seq and RNA-seq. Using ATAC-seq, we detected local changes in chromatin accessibility near the *Sd* ERV insertion, but identified little evidence for changes in chromatin accessibility elsewhere in the genome. RNA-seq, on the other hand, pointed to differential expression of 49 genes across the genome. While changes in chromatin can certainly be linked to changes in gene expression, and vice versa, the relationship is often weak or context-dependent [20], making it tricky to explicitly model the relationship between them. Nevertheless, the local change in chromatin accessibility at the *Gm13344* promoter, which is orthologous to an autoregulatory human *PTF1A* enhancer, suggests that the overexpression of *Ptf1a* in the context of the ERV insertion could be mediated at least in part by enhancer activity of the *Gm13344* promoter. Of course, our data does not prove this, and other possibilities exist, most obviously the possibility that the ERV insertion is acting as an enhancer and is solely responsible for driving the *Ptf1a* overexpression without any contribution of the *Gm13344* promoter sequence. Nevertheless, the complementary information in the ATAC-seq and RNA-seq data lead to an interesting hypothesis to be tested downstream. Additional data, such as chromatin conformation data to detect contacts between the *Gm13344* promoter and *Ptf1a* promoter, or deletion/mutation of the *Gm13344* promoter, might help clarify whether the *Gm13344* promoter is playing any role in *Ptf1a* overexpression. The ATAC-seq data also provided evidence against another hypothesis, namely that overexpression of *Ptf1a* was causing changes in chromatin accessibility at *Ptf1a* binding sites as has been observed for some other TFs [314, 43, 94]. Recent work suggests that only a minority of TFs possess the ability to impact chromatin accessibility in that way [11, 5], so it is not surprising that PTF1A cannot act in this manner. This project also demonstrated the role that omics data can play in bridging the gap between a known genetic lesion and the downstream effects, with RNA-seq data and GO enrichment analysis guiding further experimental work into a specific biological signaling pathway. PTF1A binding data, as well as data from earlier time points, could provide further information into the cascade of events – e.g., which differentially expressed genes are likely to be direct targets of PTF1A and may therefore come earlier in the cascade.

While the integration of different omics data, including public datasets, can be extremely informative, one frequent challenge is dealing with the technical heterogeneity that may be present in the data. The ataqv project (chapter III) demonstrated the substantial heterogeneity that characterizes many public ATAC-seq datasets. In addition, I analyzed original data and used QC metrics as well as the distribution of reads across the genome to quantify the impact of two technical variables on ATAC-seq results, namely the ratio of Tn5 transposase to nuclei and sequencing lane cluster density. Although a long-term biomedical research plan should be driven by meaningful biological problems, day-to-day work in a research lab is dominated by working through and around less glorious problems. This reality is typified by QC, and tools and resources that help perform QC, benchmark technical effects, and place the results into a broader context can be of great practical value. The results and resources generated by this project will hopefully assist future projects that utilize epigenomic profiling (and were put to use in the QC of all the ATAC-seq and snATAC-seq data generated for the projects in this dissertation).

In chapter IV, I performed joint analysis of snRNA-seq and snATAC-seq to map the transcriptomes and chromatin accessibility of skeletal muscle cell types, and applied the chromatin accessibility data to examine cell type GWAS SNP enrichments and nominate causal GWAS SNPs. In addition to these specific results, the downstream GWAS enrichments suggest interesting questions about cell type identity and the relative contribution of similar cell types in different tissue contexts to a given trait. For example, using UK Biobank GWAS we showed enrichment of blood and immune cell trait GWAS SNPs in the skeletal muscle immune cell ATAC-seq peaks. This is not surprising, as nominally similar cell types derived from non-muscle sources have shown similar GWAS enrichments [9, 107]. However, the extent to which the identified muscle-resident cell populations can be taken as representatives of nominally similar cell populations in other tissues remains an open question, and presumably for many traits the ‘same’ cell type in different tissues will not have equal effects on the trait of interest. These are not new topics of investigation – for example, the similarities and differences between mesenchymal stem cell populations isolated from adipose, bone marrow, and muscle has been a subject of research for years [134, 60], as has the role of tissue-resident vs. circulating immune cells [42, 272]. But high-throughput single-cell methods are moving these discussions into a new era, providing new data with which to examine old questions. The single-cell era will help clarify the extent to which we can isolate a cell type from it's tissue context, including when considering that cell type's role in traits and disease.

In chapter V, I performed the first analyses of chromatin accessibility in the F2 rats from the HCR-LCR model system. One hypothesis was that there may be chromatin accessibility differences correlating with running capacity differences between rats at rest, perhaps signaling priming of regulatory elements to become active upon a change in exercise state (enabling some rats to respond to the onset of exercise more effectively than others). However, I observed little evidence for associations between running capacity and ATAC-seq peak signal in my initial analysis. My analysis of the association between running capacity and gene expression largely recapitulated the results of previous analyses [230, 238], showing signs of signal in the data (especially in biological pathways related to fatty acid oxidation) but, in my case, no significantly associated genes. These results suggest that running capacity is a highly polygenic trait. If small effect sizes at a large number of loci drive the wide distribution of running capacity in the F2s, aggregating information across loci will be key to understanding the genetics of running capacity in this system. It is also likely that some effects will only be apparent under certain conditions; future analysis of HCR-LCR chromatin accessibility and gene expression under different exercise conditions will provide insight into the extent to which chromatin accessibility and gene expression change with exercise. In chapter V I also performed the first caQTL analysis in this rat model system, identifying 4,477 ATAC-seq peaks associated with at least one

SNP. In isolation the caQTL results provide little to no insight into the genetic factors underlying running capacity, but as genetic, epigenomic, and transcriptomic analysis of the HCR/LCR and F2 rats continues, caQTL results will be one layer of information to draw upon to help interpret future findings. For example, HCR-LCR allele frequency differences at caQTL SNPs might be one explanation for any differences in ATAC-seq peak signal we observe between the lines.

The projects carried out in this dissertation demonstrate the value as well as the challenges in applying current epigenomic and transcriptomic assays to the investigation of monogenic and polygenic traits. Continued development and refinement of experimental methods to probe molecular traits will generate exciting new datasets we can leverage to this end. Integration of this data with existing data will be key to understanding many aspects of monogenic and polygenic disease, from identifying causal mutations and genetic variants, to understanding the mechanisms through which the variants exert their effects, and the contexts in which they are relevant.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] 1000 Genomes Project Consortium, et al. (2015), A global reference for human genetic variation, *Nature*, *526*(7571), 68–74, doi:10.1038/nature15393.

[2] Abdelkhalek, H. B., et al. (2004), The mouse homeobox gene Not is required for caudal notochord development and affected by the truncate mutation., *Genes & Development*, *18*(14), 1725–1736, doi:10.1101/gad.303504.

[3] Aken, B. L., et al. (2016), The Ensembl gene annotation system., *Database*, *2016*, doi:10.1093/database/baw093.

[4] Alasoo, K., J. Rodrigues, S. Mukhopadhyay, A. J. Knights, A. L. Mann, K. Kundu, C. Hale, G. Dougan, and D. J. Gaffney (2018), Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response, *Nature Genetics*, *50*(3), 424, doi:10.1038/s41588-018-0046-7.

[5] Albanus, R. D., Y. Kyono, J. Hensley, A. Varshney, P. Orchard, J. O. Kitzman, and S. C. J. Parker (2019), Chromatin information content landscapes inform transcription factor and DNA interactions, *bioRxiv*, p. 777532, doi:10.1101/777532.

[6] Alessio, E., et al. (2019), Single cell analysis reveals the involvement of the long non-coding RNA Pvt1 in the modulation of muscle atrophy and mitochondrial network, *Nucleic Acids Research*, *47*(4), 1653–1670, doi:10.1093/nar/gkz007.

[7] Andersson, R., et al. (2014), An atlas of active enhancers across human cell types and tissues, *Nature*, *507*(7493), 455–461.

[8] Ang, S. L., and J. Rossant (1994), HNF-3 beta is essential for node and notochord formation in mouse development., *Cell*, *78*(4), 561–574, doi:10.1016/0092-8674(94)90522-3.

[9] Astle, W. J., et al. (2016), The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease, *Cell*, *167*(5), 1415–1429.e19, doi:10.1016/j.cell.2016.10.042.

[10] Aughey, G. N., A. Estacio Gomez, J. Thomson, H. Yin, and T. D. Southall (2018), CATaDa reveals global remodelling of chromatin accessibility during stem cell differentiation in vivo, *eLife*, *7*, e32,341, doi:10.7554/eLife.32341.

[11] Baek, S., I. Goldstein, and G. L. Hager (2017), Bivariate Genomic Footprinting Detects Changes in Transcription Factor Activity, *Cell Reports*, *19*(8), 1710–1722, doi:10.1016/j.celrep.2017.05.003.

[12] Baxmann, A. C., M. S. Ahmed, N. C. Marques, V. B. Menon, A. B. Pereira, G. M. Kirsztajn, and I. P. Heilberg (2008), Influence of Muscle Mass and Physical Activity on Serum and Urinary Creatinine and Serum Cystatin C, *Clinical Journal of the American Society of Nephrology : CJASN*, *3*(2), 348–354, doi:10.2215/CJN.02870707.

[13] Benjamini, Y., and Y. Hochberg (1995), Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300.

[14] Benjamini, Y., and T. P. Speed (2012), Summarizing and correcting the GC content bias in high-throughput sequencing, *Nucleic Acids Research*, *40*(10), e72, doi:10.1093/nar/gks001.

[15] Benson, G. (1999), Tandem repeats finder: a program to analyze DNA sequences, *Nucleic Acids Research*, *27*(2), 573–580, doi:10.1093/nar/27.2.573.

[16] Beres, T. M., T. Masui, G. H. Swift, L. Shi, R. M. Henke, and R. J. MacDonald (2006), PTF1 is an organ-specific and Notch-independent basic helix-loop-helix complex containing the mammalian Suppressor of Hairless (RBP-J) or its paralogue, RBP-L., *Molecular and Cellular Biology*, *26*(1), 117–130, doi:10.1128/MCB.26.1.117-130.2006.

[17] Bigot, P., et al. (2016), Functional characterization of the 12p12.1 renal cancer-susceptibility locus implicates BHLHE41, *Nature Communications*, *7*, 12,098, doi:10.1038/ncomms12098.

[18] Bossini-Castillo, L., et al. (2019), Immune disease variants modulate gene expression in regulatory CD4+ T cells and inform drug targets, *bioRxiv*, p. 654632, doi:10.1101/654632.

[19] Bouchard, C., R. Lesage, G. Lortie, J. A. Simoneau, P. Hamel, M. R. Boulay, L. Pérusse, G. Thériault, and C. Leblanc (1986), Aerobic performance in brothers, dizygotic and monozygotic twins, *Medicine and Science in Sports and Exercise*, *18*(6), 639–646.

[20] Boyle, A. P., S. Davis, H. P. Shulha, P. Meltzer, E. H. Margulies, Z. Weng, T. S. Furey, and G. E. Crawford (2008), High-resolution mapping and characterization of open chromatin across the genome., *Cell*, *132*(2), 311–322, doi:10.1016/j.cell.2007.12.014.

[21] Briscoe, J., L. Sussel, P. Serup, D. Hartigan-O'Connor, T. M. Jessell, J. L. Rubenstein, and J. Ericson (1999), Homeobox gene Nkx2.2 and specification of neuronal identity by graded Sonic hedgehog signalling., *Nature*, *398*(6728), 622–627, doi:10.1038/19315.

[22] Bronner, I. F., M. A. Quail, D. J. Turner, and H. Swerdlow (2014), Improved Protocols for Illumina Sequencing, *Current Protocols in Human Genetics*, *80*, 18.2.1–42, doi:10.1002/0471142905.hg1802s80.

[23] Buenrostro, J. D., P. G. Giresi, L. C. Zaba, H. Y. Chang, and W. J. Greenleaf (2013), Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position, *Nature Methods*, *10*(12), 1213–1218, doi:10.1038/nmeth.2688.

[24] Buenrostro, J. D., B. Wu, H. Y. Chang, and W. J. Greenleaf (2015), ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide, *Current Protocols in Molecular Biology*, *109*, 21.29.1–9, doi:10.1002/0471142727.mb2129s109.

[25] Buenrostro, J. D., B. Wu, U. M. Litzenburger, D. Ruff, M. L. Gonzales, M. P. Snyder, H. Y. Chang, and W. J. Greenleaf (2015), Single-cell chromatin accessibility reveals principles of regulatory variation, *Nature*, *523*(7561), 486–490, doi:10.1038/nature14590.

[26] Buniello, A., et al. (2019), The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019, *Nucleic Acids Research*, *47*(D1), D1005–D1012, doi:10.1093/nar/gky1120.

[27] Butler, A., P. Hoffman, P. Smibert, E. Papalexi, and R. Satija (2018), Integrating single-cell transcriptomic data across different conditions, technologies, and species, *Nature Biotechnology*, *36*(5), 411–420, doi:10.1038/nbt.4096.

[28] Calo, E., and J. Wysocka (2013), Modification of enhancer chromatin: what, how and why?, *Molecular Cell*, *49*(5), doi:10.1016/j.molcel.2013.01.038.

[29] Cannon, M. E., et al. (2019), Open Chromatin Profiling in Adipose Tissue Marks Genomic Regions with Functional Roles in Cardiometabolic Traits, *G3: Genes|Genomes|Genetics*, *9*(8), 2521–2533, doi:10.1534/g3.119.400294.

[30] Cao, J., et al. (2018), Joint profiling of chromatin accessibility and gene expression in thousands of single cells, *Science*, p. eaau0730, doi:10.1126/science.aau 0730.

[31] Casper, J., et al. (2018), The UCSC Genome Browser database: 2018 update, *Nucleic Acids Research*, *46*(D1), D762–D769, doi:10.1093/nar/gkx1020.

[32] Chakraborty, R., F. Z. Saddouk, A. C. Carrao, D. S. Krause, D. M. Greif, and K. A. Martin (2019), Promoters to Study Vascular Smooth Muscle, *Arteriosclerosis, Thrombosis, and Vascular Biology*, *39*(4), 603–612, doi:10.1161/ATVBAHA.119.312449.

[33] Chemello, F., C. Bean, P. Cancellara, P. Laveder, C. Reggiani, and G. Lanfranchi (2011), Microgenomic analysis in skeletal muscle: expression signatures of individual fast and slow myofibers, *PLoS One*, *6*(2), e16,807, doi:10.1371/journal.pone.0016807.

[34] Chemello, F., et al. (2019), Transcriptomic Analysis of Single Isolated Myofibers Identifies miR-27a-3p and miR-142-3p as Regulators of Metabolism in Skeletal Muscle, *Cell Reports*, *26*(13), 3784–3797.e8, doi:10.1016/j.celrep.2019.02.105.

[35] Chen, K.-Y., C.-H. Chiu, and L.-C. Wang (2017), Anti-apoptotic effects of Sonic hedgehog signalling through oxidative stress reduction in astrocytes co-cultured with excretory-secretory products of larval Angiostrongylus cantonensis., *Scientific Reports*, *7*, 41,574, doi:10.1038/srep41574.

[36] Chen, R., and D. K. Gifford (2017), Differential chromatin profiles partially determine transcription factor binding., *PLoS One*, *12*(7), e0179,411, doi:10.1371/journal.pone.0179411.

[37] Chen, X., U. M. Litzenburger, Y. Wei, A. N. Schep, E. L. LaGory, H. Choudhry, A. J. Giaccia, W. J. Greenleaf, and H. Y. Chang (2018), Joint single-cell DNA accessibility and protein epitope profiling reveals environmental regulation of epigenomic heterogeneity, *Nature Communications*, *9*(1), 4590, doi:10.1038/s41467-018-07115-y.

[38] Chiang, C., Y. Litingtung, E. Lee, K. E. Young, J. L. Corden, H. Westphal, and P. A. Beachy (1996), Cyclopia and defective axial patterning in mice lacking Sonic hedgehog gene function., *Nature*, *383*(6599), 407–413, doi:10.1038/383407a0.

[39] Cho, D. S., and J. D. Doles (2017), Single cell transcriptome analysis of muscle satellite cells reveals widespread transcriptional heterogeneity, *Gene*, *636*, 54–63, doi:10.1016/j.gene.2017.09.014.

[40] Choi, K.-S., C. Lee, and B. D. Harfe (2012), Sonic hedgehog in the notochord is sufficient for patterning of the intervertebral discs., *Mechanisms of Development*, *129*(9-12), 255–262, doi:10.1016/j.mod.2012.07.003.

[41] Chong, J., et al. (2015), The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities, *American Journal of Human Genetics*, *97*(2), 199–215, doi:10.1016/j.ajhg.2015.06.009.

[42] Chou, C., and M. O. Li (2018), Tissue-Resident Lymphocytes Across Innate and Adaptive Lineages, *Frontiers in Immunology*, *9*, doi:10.3389/fimmu.2018.02104.

[43] Cirillo, L. A., F. R. Lin, I. Cuesta, D. Friedman, M. Jarnik, and K. S. Zaret (2002), Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4, *Molecular Cell*, *9*(2), 279–289.

[44] Clark, S. J., et al. (2018), scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells, *Nature Communications*, *9*(1), doi:10.1038/s41467-018-03149-4.

[45] Cockell, M., B. J. Stevenson, M. Strubin, O. Hagenbüchle, and P. K. Wellauer (1989), Identification of a cell-specific DNA-binding activity that interacts with a transcriptional activator of genes expressed in the acinar pancreas., *Molecular and Cellular Biology*, *9*(6), 2464–2476.

[46] Corces, M. R., et al. (2017), An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues, *Nature Methods*, *14*(10), 959–962, doi:10.1038/nmeth.4396.

[47] Costamagna, D., H. Mommaerts, M. Sampaolesi, and P. Tylzanowski (2016), Noggin inactivation affects the number and differentiation potential of muscle progenitor cells in vivo, *Scientific Reports*, *6*(1), 1–16, doi:10.1038/srep31949.

[48] Creyghton, M. P., et al. (2010), Histone H3K27ac separates active from poised enhancers and predicts developmental state, *Proceedings of the National Academy of Sciences*, *107*(50), 21,931–21,936, doi:10.1073/pnas.1016071107.

[49] Cusanovich, D. A., R. Daza, A. Adey, H. A. Pliner, L. Christiansen, K. L. Gunderson, F. J. Steemers, C. Trapnell, and J. Shendure (2015), Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing, *Science*, *348*(6237), 910–914, doi:10.1126/science.aab1601.

[50] Davies, J. O. J., J. M. Telenius, S. J. McGowan, N. A. Roberts, S. Taylor, D. R. Higgs, and J. R. Hughes (2016), Multiplexed analysis of chromosome conformation at vastly improved sensitivity., *Nature Methods*, *13*(1), 74–80, doi:10.1038/nmeth.3664.

[51] Davis, J., L. Fresard, D. Knowles, M. Pala, C. Bustamante, A. Battle, and S. Montgomery (2016), An Efficient Multiple-Testing Adjustment for eQTL Studies that Accounts for Linkage Disequilibrium between Variants, *American Journal of Human Genetics*, *98*(1), 216–224, doi:10.1016/j.ajhg.2015.11.021.

[52] De Micheli, A. J., E. J. Laurilliard, C. L. Heinke, H. Ravichandran, P. Fraczek, S. Soueid-Baumgarten, I. De Vlaminck, O. Elemento, and B. D. Cosgrove (2020), Single-Cell Analysis of the Muscle Stem Cell Hierarchy Identifies Heterotypic Communication Signals Involved in Skeletal Muscle Regeneration, *Cell Reports*, *30*(10), 3583–3595.e5, doi:10.1016/j.celrep.2020.02.067.

[53] Delaneau, O., H. Ongen, A. A. Brown, A. Fort, N. I. Panousis, and E. T. Dermitzakis (2017), A complete tool set for molecular QTL discovery and analysis, *Nature Communications*, *8*, 15,452, doi:10.1038/ncomms15452.

[54] Dell'Orso, S., A. H. Juan, K.-D. Ko, F. Naz, J. Perovanovic, G. Gutierrez-Cruz, X. Feng, and V. Sartorelli (2019), Single cell analysis of adult mouse skeletal muscle stem cells in homeostatic and regenerative conditions, *Development*, *146*(12), doi:10.1242/dev.174177.

[55] Denas, O., R. Sandstrom, Y. Cheng, K. Beal, J. Herrero, R. C. Hardison, and J. Taylor (2015), Genome-wide comparative analysis reveals human-mouse regulatory landscape and evolution., *BMC Genomics*, *16*, 87, doi:10.1186/s12864-015-1245-6.

[56] Denker, A., and W. de Laat (2016), The second decade of 3C technologies: detailed insights into nuclear organization., *Genes & Development*, *30*(12), 1357–1382, doi:10.1101/gad.281964.116.

[57] Di Gregorio, A., R. M. Harland, M. Levine, and E. S. Casey (2002), Tail morphogenesis in the ascidian, Ciona intestinalis, requires cooperation between notochord and muscle., *Developmental Biology*, *244*(2), 385–395, doi:10.1006/dbio.2002.0582.

[58] Djebali, S., et al. (2012), Landscape of transcription in human cells., *Nature*, *489*(7414), 101–108, doi:10.1038/nature11233.

[59] Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras (2012), STAR: ultrafast universal RNA-seq aligner, *Bioinformatics*, *29*(1), 15–21.

[60] Dominici, M., et al. (2006), Minimal criteria for defining multipotent mesenchymal stromal cells. The International Society for Cellular Therapy position statement, *Cytotherapy*, *8*(4), 315–317, doi:10.1080/14653240600855905.

[61] Dunn, L. C., S. Gluecksohn-Schoenheimer, and V. Bryson (1940), A new mutation in the mouse: affecting spinal column and urogenital system, *Journal of Heredity*, *31*(8), 343–348, doi:10.1093/oxfordjournals.jhered.a104924.

[62] Echelard, Y., D. J. Epstein, B. St-Jacques, L. Shen, J. Mohler, J. A. McMahon, and A. P. McMahon (1993), Sonic hedgehog, a member of a family of putative signaling molecules, is implicated in the regulation of CNS polarity., *Cell*, *75*(7), 1417–1430, doi:10.1016/0092-8674(93)90627-3.

[63] Eisenberg, E., and E. Y. Levanon (2013), Human housekeeping genes, revisited, *Trends in Genetics*, *29*(10), 569–574, doi:10.1016/j.tig.2013.05.010.

[64] ENCODE Project Consortium (2011), A user's guide to the encyclopedia of DNA elements (ENCODE), *PLoS Biology*, *9*(4), e1001,046, doi:10.1371/journal.pbio.1001046.

[65] ENCODE Project Consortium (2012), An integrated encyclopedia of DNA elements in the human genome, *Nature*, *489*(7414), 57–74, doi:10.1038/nature11247.

[66] Epstein, D. J., A. P. McMahon, and A. L. Joyner (1999), Regionalization of Sonic hedgehog transcription along the anteroposterior axis of the mouse central nervous system is regulated by Hnf3-dependent and -independent mechanisms., *Development*, *126*(2), 281–292.

[67] Ernst, J., and M. Kellis (2012), ChromHMM: automating chromatin-state discovery and characterization, *Nature Methods*, *9*(3), 215–216, doi:10.1038/nmeth.1906.

[68] Everson, J. L., et al. (2017), Sonic hedgehog regulation of Foxf2 promotes cranial neural crest mesenchyme proliferation and is disrupted in cleft lip morphogenesis., *Development*, *144*(11), 2082–2091, doi:10.1242/dev.149930.

[69] Farahani, R. M., and M. Xaymardan (2015), Platelet-Derived Growth Factor Receptor Alpha as a Marker of Mesenchymal Stem Cells in Development and Stem Cell Biology, *Stem Cells International*, *2015*, doi:10.1155/2015/362753.

[70] Farh, K. K.-H., et al. (2015), Genetic and epigenetic fine mapping of causal autoimmune disease variants, *Nature*, *518*(7539), 337–343, doi:10.1038/nature13835.

[71] Finucane, H. K., et al. (2015), Partitioning heritability by functional annotation using genome-wide association summary statistics, *Nature Genetics*, *47*(11), 1228–1235, doi:10.1038/ng.3404.

[72] Fort, A., N. I. Panousis, M. Garieri, S. E. Antonarakis, T. Lappalainen, E. T. Dermitzakis, and O. Delaneau (2017), MBV: a method to solve sample mislabeling and detect technical bias in large combined genotype and sequencing assay datasets, *Bioinformatics*, *33*(12), 1895–1897, doi:10.1093/bioinformatics/btx074.

[73] Friedrichs, M., F. Wirsdöerfer, S. B. Flohé, S. Schneider, M. Wuelling, and A. Vortkamp (2011), BMP signaling balances proliferation and differentiation of muscle satellite cell descendants, *BMC Cell Biology*, *12*(1), 26, doi:10.1186/1471-2121-12-26.

[74] Frohman, M. A., M. K. Dush, and G. R. Martin (1988), Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer., *Proceedings of the National Academy of Sciences*, *85*(23), 8998–9002, doi:10.1073/pnas.85.23.8998.

[75] Frontera, W. R., and J. Ochala (2015), Skeletal muscle: a brief review of structure and function, *Calcified Tissue International*, *96*(3), 183–195, doi:10.1007/s00223-014-9915-y.

[76] Fu, Y., M. Sinha, C. L. Peterson, and Z. Weng (2008), The Insulator Binding Protein CTCF Positions 20 Nucleosomes around Its Binding Sites across the Human Genome, *PLoS Genetics*, *4*(7), e1000,138, doi:10.1371/journal.pgen.1000138.

[77] Fujiwara, S., S. Baek, L. Varticovski, S. Kim, and G. L. Hager (2019), High Quality ATAC-Seq Data Recovered from Cryopreserved Breast Cell Lines and Tissue, *Scientific Reports*, *9*(1), 516, doi:10.1038/s41598-018-36927-7.

[78] Fukada, S.-i., A. Uezumi, M. Ikemoto, S. Masuda, M. Segawa, N. Tanimura, H. Yamamoto, Y. Miyagoe-Suzuki, and S. Takeda (2007), Molecular signature of quiescent satellite cells in adult skeletal muscle, *Stem Cells*, *25*(10), 2448–2459, doi:10.1634/stemcells.2007-0019.

[79] Gaulton, K. J., et al. (2015), Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci, *Nature Genetics*, *47*(12), 1415–1425, doi:10.1038/ng.3437.

[80] Gazal, S., et al. (2017), Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection, *Nature Genetics*, *49*(10), 1421–1427, doi:10.1038/ng.3954.

[81] GBD 2013 Mortality and Causes of Death Collaborators (2015), Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013, *Lancet*, *385*(9963), 117–171, doi:10.1016/S0140-6736(14)61682-2.

[82] Ghandi, M., D. Lee, M. Mohammad-Noori, and M. A. Beer (2014), Enhanced regulatory sequence prediction using gapped k-mer features, *PLoS Computational Biology*, *10*(7), e1003,711, doi:10.1371/journal.pcbi.1003711.

[83] Giordani, L., et al. (2019), High-Dimensional Single-Cell Cartography Reveals Novel Skeletal Muscle-Resident Cell Populations, *Molecular Cell*, *74*(3), 609–621.e6, doi:10.1016/j.molcel.2019.02.026.

[84] Giresi, P. G., J. Kim, R. M. McDaniell, V. R. Iyer, and J. D. Lieb (2007), FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin, *Genome Research*, *17*(6), 877–885, doi:10.1101/gr.5533506.

[85] Glasgow, S. M., R. M. Henke, R. J. Macdonald, C. V. E. Wright, and J. E. Johnson (2005), Ptf1a determines GABAergic over glutamatergic neuronal cell fate in the spinal cord dorsal horn., *Development*, *132*(24), 5461–5469, doi:10.1242/dev.02167.

[86] Gluecksohn-Schoenheimer, S. (1945), The Embryonic Development of Mutants of the Sd-Strain in Mice., *Genetics*, *30*(1), 29–38.

[87] Gohl, D. M., et al. (2019), Measuring sequencer size bias using REcount: a novel method for highly accurate Illumina sequencing-based quantification, *Genome Biology*, *20*(1), 85, doi:10.1186/s13059-019-1691-6.

[88] Gomez, D., P. Swiatlowska, and G. K. Owens (2015), Epigenetic Control of Smooth Muscle Cell Identity and Lineage Memory, *Arteriosclerosis, Thrombosis, and Vascular Biology*, *35*(12), 2508–2516, doi:10.1161/ATVBAHA.115.305044.

[89] Goodwin, S., J. D. McPherson, and W. R. McCombie (2016), Coming of age: ten years of next-generation sequencing technologies, *Nature Reviews Genetics*, *17*(6), 333–351, doi:10.1038/nrg.2016.49.

[90] Grant, C. E., T. L. Bailey, and W. S. Noble (2011), FIMO: scanning for occurrences of a given motif, *Bioinformatics*, *27*(7), 1017–1018.

[91] Grüneberg, H. (1953), Genetioal studies on the skeleton of the mouse, *Journal of Genetics*, *51*(2), 317–326, doi:10.1007/BF03023300.

[92] Guo, H., P. Zhu, X. Wu, X. Li, L. Wen, and F. Tang (2013), Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing, *Genome Research*, *23*(12), 2126–2135, doi:10.1101/gr.161679.113.

[93] Ha, N.-T., S. Freytag, and H. Bickeboeller (2014), Coverage and efficiency in current SNP chips, *European Journal of Human Genetics*, *22*(9), 1124–1130, doi:10.1038/ejhg.2013.304.

[94] Hammelman, J., K. Krismer, B. Banerjee, D. K. Gifford, and R. Sherwood (2020), Identification of determinants of differential chromatin accessibility through a massively parallel genome-integrated reporter assay, *bioRxiv*, p. 2020.03.02.973396, doi:10.1101/2020.03.02.973396.

[95] Harfe, B. D., P. J. Scherz, S. Nissim, H. Tian, A. P. McMahon, and C. J. Tabin (2004), Evidence for an expansion-based temporal Shh gradient in specifying vertebrate digit identities., *Cell*, *118*(4), 517–528, doi:10.1016/j.cell.2004.07.024.

[96] Hartley, S. W., and J. C. Mullikin (2015), QoRTs: a comprehensive toolset for quality control and data processing of RNA-Seq experiments, *BMC Bioinformatics*, *16*(1), 1–7, doi:10.1186/s12859-015-0670-5.

[97] Harwood, J. C., N. A. Kent, N. D. Allen, and A. J. Harwood (2019), Nucleosome dynamics of human iPSC during neural differentiation, *EMBO Reports*, *20*(6), e46,960, doi:10.15252/embr.201846960.

[98] Heaton, H., A. M. Talman, A. Knights, M. Imaz, D. Gaffney, R. Durbin, M. Hemberg, and M. Lawniczak (2019), souporcell: Robust clustering of single cell RNAseq by genotype and ambient RNA inference without reference genotypes, *bioRxiv*, p. 699637, doi:10.1101/699637.

[99] Heid, I. M., et al. (2010), Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution, *Nature Genetics*, *42*(11), 949–960, doi:10.1038/ng.685.

[100] Heintzman, N. D., et al. (2007), Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome., *Nature Genetics*, *39*(3), 311–318, doi:10.1038/ng1966.

[101] Herrmann, B. G., S. Labeit, A. Poustka, T. R. King, and H. Lehrach (1990), Cloning of the T gene required in mesoderm formation in the mouse., *Nature*, *343*(6259), 617–622, doi:10.1038/343617a0.

[102] Hinrichs, A. S., et al. (2006), The UCSC Genome Browser Database: update 2006, *Nucleic Acids Research*, *34*(Database issue), D590–598, doi:10.1093/nar/gkj144.

[103] Hoffman, M. M., et al. (2013), Integrative annotation of chromatin elements from ENCODE data, *Nucleic Acids Research*, *41*(2), 827–841, doi:10.1093/nar/gks1284.

[104] Hormozdiari, F., et al. (2016), Colocalization of GWAS and eQTL Signals Detects Target Genes, *The American Journal of Human Genetics*, *99*(6), 1245–1260, doi:10.1016/j.ajhg.2016.10.003.

[105] Iglesias, A. I., et al. (2018), Cross-ancestry genome-wide association analysis of corneal thickness strengthens link between complex and Mendelian eye diseases, *Nature Communications*, *9*(1), 1–11, doi:10.1038/s41467-018-03646-6.

[106] Imboden, M. T., M. P. Harber, M. H. Whaley, W. H. Finch, D. L. Bishop, and L. A. Kaminsky (2018), Cardiorespiratory Fitness and Mortality in Healthy Men and Women, *Journal of the American College of Cardiology*, *72*(19), 2283–2292, doi:10.1016/j.jacc.2018.08.2166.

[107] Iotchkova, V., et al. (2019), GARFIELD classifies disease-relevant genomic features through integration of functional annotations with association signals, *Nature Genetics*, *51*(2), 343, doi:10.1038/s41588-018-0322-6.

[108] Janssen, I., S. B. Heymsfield, Z. Wang, and R. Ross (2000), Skeletal muscle mass and distribution in 468 men and women aged 18–88 yr, *Journal of Applied Physiology*, *89*(1), 81–88, doi:10.1152/jappl.2000.89.1.81.

[109] Jeong, Y., and D. J. Epstein (2003), Distinct regulators of Shh transcription in the floor plate and notochord indicate separate origins for these tissues in the mouse node., *Development*, *130*(16), 3891–3902, doi:10.1242/dev.00590.

[110] Jessell, T. M. (2000), Neuronal specification in the spinal cord: inductive signals and transcriptional codes., *Nature Reviews Genetics*, *1*(1), 20–29, doi:10.1038/35049541.

[111] Judson, R. N., R.-H. Zhang, and F. M. A. Rossi (2013), Tissue-resident mesenchymal stem/progenitor cells in skeletal muscle: collaborators or saboteurs?, *The FEBS journal*, *280*(17), 4100–4108, doi:10.1111/febs.12370.

[112] Kaestner, K. H., H. Hiemisch, B. Luckow, and G. Schutz (1994), The HNF-3 gene family of transcription factors in mice: gene structure, cDNA sequence, and mRNA distribution., *Genomics*, *20*(3), 377–385, doi:10.1006/geno.1994.1191.

[113] Kang, H. M., J. H. Sul, S. K. Service, N. A. Zaitlen, S.-Y. Kong, N. B. Freimer, C. Sabatti, and E. Eskin (2010), Variance component model to account for sample structure in genome-wide association studies, *Nature Genetics*, *42*(4), 348–354, doi:10.1038/ng.548.

[114] Kang, H. M., et al. (2018), Multiplexed droplet single-cell RNA-sequencing using natural genetic variation, *Nature Biotechnology*, *36*(1), 89–94, doi:10.1038/nbt.4042.

[115] Kaplan, N., et al. (2009), The DNA-encoded nucleosome organization of a eukaryotic genome, *Nature*, *458*(7236), 362–366, doi:10.1038/nature07667.

[116] Kashani, K., M. H. Rosner, and M. Ostermann (2020), Creatinine: From physiology to clinical application, *European Journal of Internal Medicine*, *72*, 9–14, doi:10.1016/j.ejim.2019.10.025.

[117] Kashani, K. B., E. N. Frazee, L. Kukrálová, K. Sarvottam, V. Herasevich, P. M. Young, R. Kashyap, and J. C. Lieske (2017), Evaluating Muscle Mass by Using Markers of Kidney Function: Development of the Sarcopenia Index, *Critical Care Medicine*, *45*(1), e23–e29, doi:10.1097/CCM.0000000000002013.

[118] Kawaguchi, Y., B. Cooper, M. Gannon, M. Ray, R. J. MacDonald, and C. V. E. Wright (2002), The role of the transcriptional regulator Ptf1a in converting intestinal to pancreatic progenitors., *Nature Genetics*, *32*(1), 128–134, doi:10.1038/ng959.

[119] Kaya-Okur, H. S., S. J. Wu, C. A. Codomo, E. S. Pledger, T. D. Bryson, J. G. Henikoff, K. Ahmad, and S. Henikoff (2019), CUT&Tag for efficient epigenomic profiling of small samples and single cells, *bioRxiv*, p. 568915, doi:10.1101/568915.

[120] Keegan, C. E., J. E. Hutz, T. Else, M. Adamska, S. P. Shah, A. E. Kent, J. M. Howes, W. G. Beamer, and G. D. Hammer (2005), Urogenital and caudal dysgenesis in adrenocortical dysplasia (acd) mice is caused by a splicing mutation in a novel telomeric regulator., *Human Molecular Genetics*, *14*(1), 113–123, doi:10.1093/hmg/ddi011.

[121] Kelly, R. D. W., A. Mahmud, M. McKenzie, I. A. Trounce, and J. C. St John (2012), Mitochondrial DNA copy number is regulated in a tissue specific manner by DNA methylation of the nuclear-encoded DNA polymerase gamma A, *Nucleic Acids Research*, *40*(20), 10,124–10,138, doi:10.1093/nar/gks770.

[122] Kent, W. J., C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler (2002), The human genome browser at UCSC, *Genome Research*, *12*(6), 996–1006.

[123] Kent, W. J., A. S. Zweig, G. Barber, A. S. Hinrichs, and D. Karolchik (2010), BigWig and BigBed: enabling browsing of large distributed datasets, *Bioinformatics*, *26*(17), 2204–2207, doi:10.1093/bioinformatics/btq351.

[124] Kharchenko, P. V., M. Y. Tolstorukov, and P. J. Park (2008), Design and analysis of ChIP-seq experiments for DNA-binding proteins., *Nature Biotechnology*, *26*(12), 1351–1359, doi:10.1038/nbt.1508.

[125] Kheradpour, P., and M. Kellis (2014), Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments, *Nucleic Acids Research*, *42*(5), 2976–2987, doi:10.1093/nar/gkt1249.

[126] Khetan, S., R. Kursawe, A. Youn, N. Lawlor, A. Jillette, E. J. Marquez, D. Ucar, and M. L. Stitzel (2018), Type 2 Diabetes–Associated Genetic Variants Regulate Chromatin Accessibility in Human Islets, *Diabetes*, *67*(11), 2466–2477, doi:10.2337/db18-0393.

[127] Kim, P. C., R. Mo, and C. Hui Cc (2001), Murine models of VACTERL syndrome: Role of sonic hedgehog signaling pathway., *Journal of Pediatric Surgery*, *36*(2), 381–384, doi:10.1053/jpsu.2001.20722.

[128] Kimmel, S. G., R. Mo, C. C. Hui, and P. C. Kim (2000), New mouse models of congenital anorectal malformations., *Journal of Pediatric Surgery*, *35*(2), 227–30; discussion 230–231, doi:10.1016/s0022-3468(00)90014-9.

[129] Klimczak, A., U. Kozlowska, and M. Kurpisz (2018), Muscle Stem/Progenitor Cells and Mesenchymal Stem Cells of Bone Marrow Origin for Skeletal Muscle Regeneration in Muscular Dystrophies, *Archivum Immunologiae et Therapiae Experimentalis*, *66*(5), 341–354, doi:10.1007/s00005-018-0509-7.

[130] Koch, L. G., and S. L. Britton (2001), Artificial selection for intrinsic aerobic endurance running capacity in rats, *Physiological Genomics*, *5*(1), 45–52, doi:10.1152/physiolgenomics.2001.5.1.45.

[131] Koch, L. G., and S. L. Britton (2005), Divergent selection for aerobic capacity in rats as a model for complex disease, *Integrative and Comparative Biology*, *45*(3), 405–415, doi:10.1093/icb/45.3.405.

[132] Koch, L. G., S. L. Britton, and U. Wisløff (2012), A Rat Model System to Study Complex Disease Risks, Fitness, Aging, and Longevity, *Trends in cardiovascular medicine*, *22*(2), 29–34, doi:10.1016/j.tcm.2012.06.007.

[133] Koch, L. G., et al. (2011), Intrinsic aerobic capacity sets a divide for aging and longevity, *Circulation Research*, *109*(10), 1162–1172, doi:10.1161/CIRCRESAHA.111.253807.

[134] Kozlowska, U., A. Krawczenko, K. Futoma, T. Jurek, M. Rorat, D. Patrzalek, and A. Klimczak (2019), Similarities and differences between mesenchymal stem/progenitor cells derived from various human tissues, *World Journal of Stem Cells*, *11*(6), 347–374, doi:10.4252/wjsc.v11.i6.347.

[135] Krapp, A., M. Knofler, S. Frutiger, G. J. Hughes, O. Hagenbuchle, and P. K. Wellauer (1996), The p48 DNA-binding subunit of transcription factor PTF1 is a new exocrine pancreas-specific basic helix-loop-helix protein., *The EMBO journal*, *15*(16), 4317–4329.

[136] Krapp, A., M. Knofler, B. Ledermann, K. Burki, C. Berney, N. Zoerkler, O. Hagenbuchle, and P. K. Wellauer (1998), The bHLH protein PTF1-p48 is essential for the formation of the exocrine and the correct spatial organization of the endocrine pancreas., *Genes & Development*, *12*(23), 3752–3763, doi:10.1101/gad.12.23.3752.

[137] Kristiansen, M., J. H. Graversen, C. Jacobsen, O. Sonne, H. J. Hoffman, S. K. Law, and S. K. Moestrup (2001), Identification of the haemoglobin scavenger receptor, *Nature*, *409*(6817), 198–201, doi:10.1038/35051594.

[138] Kumasaka, N., A. J. Knights, and D. J. Gaffney (2016), Fine-mapping cellular QTLs with RASQUAL and ATAC-seq, *Nature Genetics*, *48*(2), 206–213, doi:10.1038/ng.3467.

[139] Kumasaka, N., A. J. Knights, and D. J. Gaffney (2019), High-resolution genetic mapping of putative causal interactions between regions of open chromatin, *Nature Genetics*, *51*(1), 128, doi:10.1038/s41588-018-0278-6.

[140] Kundaje, A., et al. (2015), Integrative analysis of 111 reference human epigenomes., *Nature*, *518*(7539), 317–330, doi:10.1038/nature14248.

[141] Lamparter, D., D. Marbach, R. Rueedi, S. Bergmann, and Z. Kutalik (2017), Genome-wide association between transcription factor expression and chromatin accessibility reveals regulators of chromatin accessibility, *PLoS Computational Biology*, *13*(1), e1005,311.

[142] Lander, E. S., et al. (2001), Initial sequencing and analysis of the human genome, *Nature*, *409*(6822), 860–921, doi:10.1038/35057062.

[143] Landt, S. G., et al. (2012), ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia, *Genome Research*, *22*(9), 1813–1831, doi:10.1101/gr.136184.111.

[144] Landt, S. G., et al. (2012), ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia., *Genome Research*, *22*(9), 1813–1831, doi:10.1101/gr.136184.111.

[145] Latos, P. A., et al. (2012), Airn transcriptional overlap, but not its lncRNA products, induces imprinted Igf2r silencing., *Science*, *338*(6113), 1469–1472, doi:10.1126/science.1228110.

[146] Lau, S. K., P. G. Chu, and L. M. Weiss (2004), CD163: a specific marker of macrophages in paraffin-embedded tissue samples, *American Journal of Clinical Pathology*, *122*(5), 794–801, doi:10.1309/QHD6-YFN8-1KQX-UUH6.

[147] Laukkanen, J. A., T. A. Lakka, R. Rauramaa, R. Kuhanen, J. M. Venäläinen, R. Salonen, and J. T. Salonen (2001), Cardiovascular fitness as a predictor of mortality in men, *Archives of Internal Medicine*, *161*(6), 825–831, doi:10.1001/archinte.161.6.825.

[148] Lee, C., S. Patil, and M. A. Sartor (2016), RNA-Enrich: a cut-off free functional enrichment testing method for RNA-seq with improved detection power, *Bioinformatics*, *32*(7), 1100–1102, doi:10.1093/bioinformatics/btv694.

[149] Lee, C.-K., Y. Shibata, B. Rao, B. D. Strahl, and J. D. Lieb (2004), Evidence for nucleosome depletion at active regulatory regions genome-wide, *Nature Genetics*, *36*(8), 900–905, doi:10.1038/ng1400.

[150] Lee, D. (2016), LS-GKM: a new gkm-SVM for large-scale datasets, *Bioinformatics*, *32*(14), 2196–2198, doi:10.1093/bioinformatics/btw142.

[151] Lee, D., D. U. Gorkin, M. Baker, B. J. Strober, A. L. Asoni, A. S. McCallion, and M. A. Beer (2015), A method to predict the impact of regulatory variants from DNA sequence, *Nature Publishing Group*, *47*(8), 955–961.

[152] Lee, H., et al. (2020), Diagnostic utility of transcriptome sequencing for rare Mendelian diseases, *Genetics in Medicine*, *22*(3), 490–499, doi:10.1038/s41436-019-0672-1.

[153] Leslie, E. J., et al. (2015), Identification of functional variants for cleft lip with or without cleft palate in or near PAX7, FGFR2, and NOG by targeted sequencing of GWAS loci, *American Journal of Human Genetics*, *96*(3), 397–411, doi:10.1016/j.ajhg.2015.01.004.

[154] Levey, A. S., J. P. Bosch, J. B. Lewis, T. Greene, N. Rogers, and D. Roth (1999), A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation. Modification of Diet in Renal Disease Study Group, *Annals of Internal Medicine*, *130*(6), 461–470, doi:10.7326/0003-4819-130-6-199903160-00002.

[155] Le Gros, M., et al. (2016), Soft X-Ray Tomography Reveals Gradual Chromatin Compaction and Reorganization during Neurogenesis In Vivo, *Cell Reports*, *17*(8), 2125–2136, doi:10.1016/j.celrep.2016.10.060.

[156] Li, H., and R. Durbin (2009), Fast and accurate short read alignment with Burrows-Wheeler transform, *Bioinformatics*, *25*(14), 1754–1760, doi:10.1093/bioinformatics/btp324.

[157] Li, H., et al. (2009), The Sequence Alignment/Map format and SAMtools, *Bioinformatics*, *25*(16), 2078–2079, doi:10.1093/bioinformatics/btp352.

[158] Li, Q., J. B. Brown, H. Huang, and P. J. Bickel (2011), Measuring reproducibility of high-throughput experiments, *The Annals of Applied Statistics*, *5*(3), 1752–1779, doi:10.1214/11-AOAS466.

[159] Li, Z., M. H. Schulz, T. Look, M. Begemann, M. Zenke, and I. G. Costa (2019), Identification of transcription factor binding sites using ATAC-seq, *Genome Biology*, *20*(1), 45, doi:10.1186/s13059-019-1642-2.

[160] Lister, R., R. C. O'Malley, J. Tonti-Filippini, B. D. Gregory, C. C. Berry, A. H. Millar, and J. R. Ecker (2008), Highly integrated single-base resolution maps of the epigenome in Arabidopsis, *Cell*, *133*(3), 523–536, doi:10.1016/j.cell.2008.03.029.

[161] Litingtung, Y., and C. Chiang (2000), Specification of ventral neuron types is mediated by an antagonistic interaction between Shh and Gli3., *Nature Neuroscience*, *3*(10), 979–985, doi:10.1038/79916.

[162] Liu, C.-T., et al. (2014), Multi-ethnic fine-mapping of 14 central adiposity loci, *Human Molecular Genetics*, *23*(17), 4738–4744, doi:10.1093/hmg/ddu183.

[163] Liu, Y.-x., et al. (2019), Dissecting cell diversity and connectivity in skeletal muscle for myogenesis, *Cell Death & Disease*, *10*(6), doi:10.1038/s41419-019-1647-5.

[164] Lorch, Y., J. W. LaPointe, and R. D. Kornberg (1987), Nucleosomes inhibit the initiation of transcription but allow chain elongation with the displacement of histones, *Cell*, *49*(2), 203–210, doi:10.1016/0092-8674(87)90561-7.

[165] Love, M. I., W. Huber, and S. Anders (2014), Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, *Genome Biology*, *15*(12), 31–21.

[166] Lugani, F., et al. (2013), A Retrotransposon Insertion in the 5' Regulatory Domain of Ptf1a Results in Ectopic Gene Expression and Multiple Congenital Defects in Danforth's Short Tail Mouse, *PLoS Genetics*, *9*(2), e1003,206, doi:10.1371/journal.pgen.1003206.

[167] Luger, K., A. W. Mäder, R. K. Richmond, D. F. Sargent, and T. J. Richmond (1997), Crystal structure of the nucleosome core particle at 2.8 A resolution, *Nature*, *389*(6648), 251–260, doi:10.1038/38444.

[168] Lupiáñez, D. G., M. Spielmann, and S. Mundlos (2016), Breaking TADs: How Alterations of Chromatin Domains Result in Disease, *Trends in Genetics*, *32*(4), 225–237.

[169] Lupiáñez, D. G., et al. (2015), Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions, *Cell*, *161*(5), 1012–1025, doi:10.1016/j.cell.2015.04.004.

[170] Lähnemann, D., et al. (2020), Eleven grand challenges in single-cell data science, *Genome Biology*, *21*(1), 31, doi:10.1186/s13059-020-1926-6.

[171] Mahajan, A., et al. (2014), Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility, *Nature Genetics*, *46*(3), 234–244, doi:10.1038/ng.2897.

[172] Mahajan, A., et al. (2018), Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps, *Nature Genetics*, *50*(11), 1505, doi:10.1038/s41588-018-0241-6.

[173] Maier, J. A., Y. Lo, and B. D. Harfe (2013), Foxa1 and Foxa2 are required for formation of the intervertebral discs., *PLoS One*, *8*(1), e55,528, doi:10.1371/journal.pone.0055528.

[174] Mandsager, K., S. Harb, P. Cremer, D. Phelan, S. E. Nissen, and W. Jaber (2018), Association of Cardiorespiratory Fitness With Long-term Mortality Among Adults Undergoing Exercise Treadmill Testing, *JAMA Network Open*, *1*(6), e183,605, doi:10.1001/jamanetworkopen.2018.3605.

[175] Manning, A. K., et al. (2012), A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance, *Nature Genetics*, *44*(6), 659–669, doi:10.1038/ng.2274.

[176] Martin, M. (2011), Cutadapt removes adapter sequences from high-throughput sequencing reads, *EMBnet.journal*, *17*(1), 10–12, doi:10.14806/ej.17.1.200, number: 1.

[177] Masui, T., G. H. Swift, M. A. Hale, D. M. Meredith, J. E. Johnson, and R. J. MacDonald (2008), Transcriptional Autoregulation Controls Pancreatic Ptf1a Expression during Development and Adulthood, *Molecular and Cellular Biology*, *28*(17), 5458–5468.

[178] Mattis, K. K., and A. L. Gloyn (2020), From Genetic Association to Molecular Mechanisms for Islet-cell Dysfunction in Type 2 Diabetes, *Journal of Molecular Biology*, *432*(5), 1551–1578, doi:10.1016/j.jmb.2019.12.045.

[179] Maurano, M. T., et al. (2012), Systematic Localization of Common Disease-Associated Variation in Regulatory DNA, *Science*, *337*(6099), 1190–1195, doi:10.1126/science.1222794.

[180] McCann, M. R., O. J. Tamplin, J. Rossant, and C. A. Seguin (2012), Tracing notochord-derived cells using a Noto-cre mouse: implications for intervertebral disc development., *Disease Models & Mechanisms*, *5*(1), 73–82, doi:10.1242/dmm.008128.

[181] McGinnis, C. S., L. M. Murrow, and Z. J. Gartner (2019), DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors, *Cell Systems*, *8*(4), 329–337.e4, doi:10.1016/j.cels.2019.03.003.

[182] Meredith, D. M., T. Masui, G. H. Swift, R. J. MacDonald, and J. E. Johnson (2009), Multiple transcriptional mechanisms control Ptf1a levels during neural development including autoregulation by the PTF1-J complex., *The Journal of Neuroscience*, *29*(36), 11,139–11,148, doi:10.1523/JNEUROSCI.2303-09.2009.

[183] Meredith, D. M., et al. (2013), Program Specificity for Ptf1a in Pancreas versus Neural Tube Development Correlates with Distinct Collaborating Cofactors and Chromatin Accessibility, *Molecular and Cellular Biology*, *33*(16), 3166–3179, doi:10.1128/MCB.00364-13.

[184] Meuleman, W., et al. (2019), Index and biological spectrum of accessible DNA elements in the human genome, *bioRxiv*, p. 822510, doi:10.1101/822510.

[185] Micheli, A. J. D., J. A. Spector, O. Elemento, and B. D. Cosgrove (2020), A reference single-cell transcriptomic atlas of human skeletal muscle tissue reveals bifurcated muscle stem cell populations, *bioRxiv*, p. 2020.01.21.914713, doi: 10.1101/2020.01.21.914713.

[186] Mo, R., J. H. Kim, J. Zhang, C. Chiang, C. C. Hui, and P. C. Kim (2001), Anorectal malformations caused by defects in sonic hedgehog signaling., *The American Journal of Pathology*, *159*(2), 765–774, doi:10.1016/S0002-9440(10)61747-6.

[187] Mona, B., J. M. Avila, D. M. Meredith, R. K. Kollipara, and J. E. Johnson (2016), Regulating the dorsal neural tube expression of Ptf1a through a distal 3' enhancer., *Developmental Biology*, *418*(1), 216–225, doi:10.1016/j.ydbio.2016.06.033.

[188] Monaghan, A. P., K. H. Kaestner, E. Grau, and G. Schutz (1993), Postimplantation expression patterns indicate a role for the mouse forkhead/HNF-3 alpha, beta and gamma genes in determination of the definitive endoderm, chordamesoderm and neuroectoderm., *Development*, *119*(3), 567–578.

[189] Montefiori, L., L. Hernandez, Z. Zhang, Y. Gilad, C. Ober, G. Crawford, M. Nobrega, and N. Jo Sakabe (2017), Reducing mitochondrial reads in ATAC-seq using CRISPR/Cas9, *Scientific Reports*, *7*(1), 2451, doi:10.1038/s41598-017-02547-w.

[190] Morris, E. M., et al. (2019), Intrinsic High Aerobic Capacity in Male Rats Protects Against Diet-Induced Insulin Resistance, *Endocrinology*, *160*(5), 1179–1192, doi:10.1210/en.2019-00118.

[191] Mortazavi, A., B. A. Williams, K. McCue, L. Schaeffer, and B. Wold (2008), Mapping and quantifying mammalian transcriptomes by RNA-Seq, *Nature Methods*, *5*(7), 621–628, doi:10.1038/nmeth.1226.

[192] Mouse Genome Sequencing Consortium, et al. (2002), Initial sequencing and comparative analysis of the mouse genome, *Nature*, *420*(6915), 520–562, doi: 10.1038/nature01262.

[193] Nagalakshmi, U., Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder (2008), The transcriptional landscape of the yeast genome defined by RNA sequencing, *Science*, *320*(5881), 1344–1349, doi:10.1126/science.1158441.

[194] Nagano, T., Y. Lubling, T. J. Stevens, S. Schoenfelder, E. Yaffe, W. Dean, E. D. Laue, A. Tanay, and P. Fraser (2013), Single-cell Hi-C reveals cell-to-cell variability in chromosome structure, *Nature*, *502*(7469), 59–64, doi:10.1038/nature12593.

[195] Naumann, S., D. Reutzel, M. Speicher, and H.-J. Decker (2001), Complete karyotype characterization of the K562 cell line by combined application of G-banding, multiplex-fluorescence in situ hybridization, fluorescence in situ hybridization, and comparative genomic hybridization, *Leukemia Research*, *25*(4), 313–322, doi:10.1016/S0145-2126(00)00125-9.

[196] Nishida, K., M. Hoshino, Y. Kawaguchi, and F. Murakami (2010), Ptf1a directly controls expression of immunoglobulin superfamily molecules Nephrin and Neph3 in the developing central nervous system., *The Journal of Biological Chemistry*, *285*(1), 373–380, doi:10.1074/jbc.M109.060657.

[197] Noguchi, K. K., O. H. Cabrera, B. S. Swiney, P. Salinas-Contreras, J. K. Smith, and N. B. Farber (2015), Hedgehog regulates cerebellar progenitor cell and medulloblastoma apoptosis., *Neurobiology of Disease*, *83*, 35–43, doi:10.1016/j.nbd.2015.08.020.

[198] Obata, J., M. Yano, H. Mimura, T. Goto, R. Nakayama, Y. Mibu, C. Oka, and M. Kawaichi (2001), p48 subunit of mouse PTF1 binds to RBP-Jkappa/CBF-1, the intracellular mediator of Notch signalling, and is expressed in the neural tube of early stage embryos., *Genes to Cells*, *6*(4), 345–360, doi:10.1046/j.1365-2443.2001.00422.x.

[199] O'Donnell, M., C.-S. Hong, X. Huang, R. J. Delnicki, and J.-P. Saint-Jeannet (2006), Functional analysis of Sox8 during neural crest development in Xenopus., *Development*, *133*(19), 3817–3826, doi:10.1242/dev.02558.

[200] Oprescu, S. N., F. Yue, J. Qiu, L. F. Brito, and S. Kuang (2020), Temporal Dynamics and Heterogeneity of Cell Populations during Skeletal Muscle Regeneration, *iScience*, *23*(4), doi:10.1016/j.isci.2020.100993.

[201] Orchard, P., J. S. White, P. E. Thomas, A. Mychalowych, A. Kiseleva, J. Hensley, B. Allen, S. C. J. Parker, and C. E. Keegan (2019), Genome-wide chromatin accessibility and transcriptome profiling show minimal epigenome changes and coordinated transcriptional dysregulation of hedgehog signaling in Danforth's short tail mice, *Human Molecular Genetics*, *28*(5), 736–750, doi:10.1093/hmg/ddy378.

[202] Orchard, P., Y. Kyono, J. Hensley, J. O. Kitzman, and S. C. J. Parker (2020), Quantification, Dynamic Visualization, and Validation of Bias in

ATAC-Seq Data with ataqv, *Cell Systems*, *10*(3), 298–306.e4, doi:10.1016/j.cels.2020.02.009.

[203] Ou, J., H. Liu, J. Yu, M. A. Kelliher, L. H. Castilla, N. D. Lawson, and L. J. Zhu (2018), ATACseqQC: a Bioconductor package for post-alignment quality assessment of ATAC-seq data, *BMC Genomics*, *19*(1), 169, doi:10.1186/s12864-018-4559-3.

[204] Overmyer, K., et al. (2015), Maximal Oxidative Capacity during Exercise Is Associated with Skeletal Muscle Fuel Selection and Dynamic Changes in Mitochondrial Protein Acetylation, *Cell Metabolism*, *21*(3), 468–478, doi:10.1016/j.cmet.2015.02.007.

[205] Pabst, O., H. Herbrand, and H. H. Arnold (1998), Nkx2-9 is a novel homeobox transcription factor which demarcates ventral domains in the developing mouse CNS., *Mechanisms of Development*, *73*(1), 85–93, doi:10.1016/s0925-4773(98)00035-5.

[206] Pabst, O., H. Herbrand, N. Takuma, and H. H. Arnold (2000), NKX2 gene expression in neuroectoderm but not in mesendodermally derived structures depends on sonic hedgehog in mouse embryos., *Development Genes and Evolution*, *210*(1), 47–50, doi:10.1007/pl00008188.

[207] Paralkar, V. R., et al. (2016), Unlinking an lncRNA from Its Associated cis Element., *Molecular Cell*, *62*(1), 104–110, doi:10.1016/j.molcel.2016.02.029.

[208] Parker, S. C. J., et al. (2013), Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants, *Proceedings of the National Academy of Sciences*, *110*(44), 17,921–17,926, doi:10.1073/pnas.1317023110.

[209] Pasquali, L., et al. (2014), Pancreatic islet enhancer clusters enriched in type 2 diabetes risk–associated variants, *Nature Genetics*, *46*(2), 136–143, doi:10.1038/ng.2870.

[210] Patel, N., et al. (2017), A novel mechanism for variable phenotypic expressivity in Mendelian diseases uncovered by an AU-rich element (ARE)-creating mutation, *Genome Biology*, *18*, doi:10.1186/s13059-017-1274-3.

[211] Pauli, R. M. (1994), Lower mesodermal defects: a common cause of fetal and early neonatal death, *American Journal of Medical Genetics*, *50*(2), 154–172, doi:10.1002/ajmg.1320500206.

[212] Pawlikowski, B., N. D. Betta, T. Elston, R. O'Rourke, K. Jones, and B. B. Olwin (2019), A cellular atlas of skeletal muscle regeneration and aging, *bioRxiv*, p. 635805, doi:10.1101/635805.

[213] Pennimpede, T., J. Proske, A. Konig, J. A. Vidigal, M. Morkel, J. B. Bramsen, B. G. Herrmann, and L. Wittler (2012), In vivo knockdown of Brachyury results in skeletal defects and urorectal malformations resembling caudal regression syndrome., *Developmental Biology*, *372*(1), 55–67, doi:10.1016/j.ydbio.2012.09.003.

[214] Picelli, S., K. Björklund, B. Reinius, S. Sagasser, G. Winberg, and R. Sandberg (2014), Tn5 transposase and tagmentation procedures for massively scaled sequencing projects, *Genome Research*, *24*(12), 2033–2040, doi:10.1101/gr.177881.114.

[215] Pillon, N. J., P. J. Bilan, L. N. Fink, and A. Klip (2012), Cross-talk between skeletal muscle and immune cells: muscle-derived mediators and metabolic implications, *American Journal of Physiology-Endocrinology and Metabolism*, *304*(5), E453–E465, doi:10.1152/ajpendo.00553.2012.

[216] Pittenger, M. F., D. E. Discher, B. M. Péault, D. G. Phinney, J. M. Hare, and A. I. Caplan (2019), Mesenchymal stem cell perspective: cell biology to clinical progress, *npj Regenerative Medicine*, *4*(1), 1–15, doi:10.1038/s41536-019-0083-6.

[217] Plouhinec, J.-L., C. Granier, C. Le Mentec, K. A. Lawson, D. Saberan-Djoneidi, J. Aghion, D. L. Shi, J. Collignon, and S. Mazan (2004), Identification of the mammalian Not gene via a phylogenomic approach., *Gene expression patterns : GEP*, *5*(1), 11–22, doi:10.1016/j.modgep.2004.06.010.

[218] Polubriaginof, F. C. G., et al. (2018), Disease Heritability Inferred from Familial Relationships Reported in Medical Records, *Cell*, *173*(7), 1692–1704.e11, doi:10.1016/j.cell.2018.04.032.

[219] Preissl, S., et al. (2018), Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation, *Nature Neuroscience*, *21*(3), 432–439, doi:10.1038/s41593-018-0079-3.

[220] Prince, M. J., F. Wu, Y. Guo, L. M. Gutierrez Robledo, M. O'Donnell, R. Sullivan, and S. Yusuf (2015), The burden of disease in older people and implications for health policy and practice, *Lancet*, *385*(9967), 549–562, doi:10.1016/S0140-6736(14)61347-7.

[221] Pruim, R. J., R. P. Welch, S. Sanna, T. M. Teslovich, P. S. Chines, T. P. Gliedt, M. Boehnke, G. R. Abecasis, and C. J. Willer (2010), LocusZoom: regional visualization of genome-wide association scan results, *Bioinformatics*, *26*(18), 2336–2337, doi:10.1093/bioinformatics/btq419.

[222] Qiu, K., D. Xu, L. Wang, X. Zhang, N. Jiao, L. Gong, and J. Yin (2020), Association Analysis of Single-Cell RNA Sequencing and Proteomics Reveals a Vital Role of Ca2+ Signaling in the Determination of Skeletal Muscle Development Potential, *Cells*, *9*(4), doi:10.3390/cells9041045.

[223] Quach, B., and T. S. Furey (2016), DeFCoM: analysis and modeling of transcription factor binding sites using a motif-centric genomic footprinter, *Bioinformatics*, p. btw740, doi:10.1093/bioinformatics/btw740.

[224] Quang, D. X., M. R. Erdos, S. C. J. Parker, and F. S. Collins (2015), Motif signatures in stretch enhancers are enriched for disease-associated genetic variants, *Epigenetics & Chromatin*, *8*, 23, doi:10.1186/s13072-015-0015-7.

[225] Quinlan, A. R. (2014), BEDTools: The Swiss-Army Tool for Genome Feature Analysis, *Current Protocols in Bioinformatics*, *47*(1), 11.12.1–11.12.34, doi:10.1002/0471250953.bi1112s47.

[226] Quinlan, A. R. (2014), BEDTools: The Swiss-Army Tool for Genome Feature Analysis., *Current Protocols in Bioinformatics*, *47*, 11.12.1–34, doi:10.1002/0471250953.bi1112s47.

[227] Rada-Iglesias, A., R. Bajpai, T. Swigut, S. A. Brugmann, R. A. Flynn, and J. Wysocka (2011), A unique chromatin signature uncovers early developmental enhancers in humans, *Nature*, *470*(7333), 279–283, doi:10.1038/nature09692.

[228] Rai, V., et al. (2020), Single-cell ATAC-Seq in human pancreatic islets and deep learning upscaling of rare cells reveals cell-specific type 2 diabetes regulatory signatures, *Molecular Metabolism*, *32*, 109–121, doi:10.1016/j.molmet.2019.12.006.

[229] Ramani, V., X. Deng, R. Qiu, K. L. Gunderson, F. J. Steemers, C. M. Disteche, W. S. Noble, Z. Duan, and J. Shendure (2017), Massively multiplex single-cell Hi-C, *Nature Methods*, *14*(3), 263–266, doi:10.1038/nmeth.4155.

[230] Ramdas, S. (2018), Genomics of Complex Traits: Methods and Applications, Ph.D. thesis, University of Michigan.

[231] Ramdas, S., A. B. Ozel, M. K. Treutelaar, K. Holl, M. Mandel, L. C. S. Woods, and J. Z. Li (2019), Extended regions of suspected mis-assembly in the rat reference genome, *Scientific Data*, *6*(1), 39, doi:10.1038/s41597-019-0041-6.

[232] Raue, U., T. A. Trappe, S. T. Estrem, H.-R. Qian, L. M. Helvering, R. C. Smith, and S. Trappe (2012), Transcriptome signature of resistance exercise adaptations: mixed muscle and fiber type specific profiles in young and old adults, *Journal of Applied Physiology*, *112*(10), 1625–1636, doi:10.1152/japplphysiol.00435.2011.

[233] Rausch, T., M. Hsi-Yang Fritz, J. O. Korbel, and V. Benes (2019), Alfred: interactive multi-sample BAM alignment statistics, feature counting and feature annotation for long- and short-read sequencing, *Bioinformatics*, *35*(14), 2489–2491, doi:10.1093/bioinformatics/bty1007.

[234] Rebuzzini, P., T. Neri, G. Mazzini, M. Zuccotti, C. A. Redi, and S. Garagna (2008), Karyotype analysis of the euploid cell population of a mouse embryonic

stem cell line revealed a high incidence of chromosome abnormalities that varied during culture, *Cytogenetic and Genome Research*, *121*(1), 18–24, doi:10.1159/000124377.

[235] Reimann, J., K. Brimah, R. Schröder, A. Wernig, J. R. Beauchamp, and T. A. Partridge (2004), Pax7 distribution in human skeletal muscle biopsies and myogenic tissue cultures, *Cell and Tissue Research*, *315*(2), 233–242, doi:10.1007/s00441-003-0833-y.

[236] Relaix, F., and P. S. Zammit (2012), Satellite cells are essential for skeletal muscle regeneration: the cell on the edge returns centre stage, *Development*, *139*(16), 2845–2856, doi:10.1242/dev.069088.

[237] Ren, Y.-y., L. G. Koch, S. L. Britton, N. R. Qi, M. K. Treutelaar, C. F. Burant, and J. Z. Li (2015), High-density SNP array and genome sequencing reveal signatures of selection in a divergent selection rat model for aerobic running capacity, *bioRxiv*, p. 032441.

[238] Ren, Y.-y., L. G. Koch, S. L. Britton, N. R. Qi, M. K. Treutelaar, C. F. Burant, and J. Z. Li (2016), Selection-, age-, and exercise-dependence of skeletal muscle gene expression patterns in a rat model of metabolic fitness., *Physiological Genomics*, *48*(11), 816–825.

[239] Ren, Y.-y., et al. (2013), Genetic Analysis of a Rat Model of Aerobic Capacity and Metabolic Fitness, *PLoS One*, *8*(10), e77,588, doi:10.1371/journal.pone.0077588.

[240] Ribes, V., et al. (2010), Distinct Sonic Hedgehog signaling dynamics specify floor plate and ventral neuronal progenitors in the vertebrate neural tube., *Genes & Development*, *24*(11), 1186–1200, doi:10.1101/gad.559910.

[241] Richards, J. B., et al. (2009), A Genome-Wide Association Study Reveals Variants in ARL15 that Influence Adiponectin Levels, *PLoS Genetics*, *5*(12), doi:10.1371/journal.pgen.1000768.

[242] Ristola, M., and S. Lehtonen (2014), Functions of the podocyte proteins nephrin and Neph3 and the transcriptional regulation of their genes., *Clinical Science*, *126*(5), 315–328, doi:10.1042/CS20130258.

[243] Roadmap Epigenomics Consortium, et al. (2015), Integrative analysis of 111 reference human epigenomes, *Nature*, *518*(7539), 317–330, doi:10.1038/nature14248.

[244] Rohland, N., and D. Reich (2012), Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture, *Genome Research*, *22*(5), 939–946, doi:10.1101/gr.128124.111.

[245] Rotem, A., O. Ram, N. Shoresh, R. A. Sperling, A. Goren, D. A. Weitz, and B. E. Bernstein (2015), Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state, *Nature Biotechnology*, *33*(11), 1165–1172, doi:10.1038/nbt.3383.

[246] Rubenstein, A. B., et al. (2020), Single-cell transcriptional profiles in human skeletal muscle, *Scientific Reports*, *10*(1), 1–15, doi:10.1038/s41598-019-57110-6.

[247] Runck, L. A., et al. (2014), Defining the molecular pathologies in cloaca malformation: similarities between mouse and human., *Disease Models & Mechanisms*, *7*(4), 483–493, doi:10.1242/dmm.014530.

[248] Sadeh, R., and C. D. Allis (2011), Genome-wide "Re"-Modeling of Nucleosome Positions, *Cell*, *147*(2), 263–266, doi:10.1016/j.cell.2011.09.042.

[249] Sadler, J. E. (1998), Biochemistry and genetics of von Willebrand factor, *Annual Review of Biochemistry*, *67*, 395–424, doi:10.1146/annurev.biochem.67.1.395.

[250] Sartori, R., P. Gregorevic, and M. Sandri (2014), TGFB and BMP signaling in skeletal muscle: potential significance for muscle-related disease, *Trends in Endocrinology and Metabolism*, *25*(9), 464–471, doi:10.1016/j.tem.2014.06.002.

[251] Sartori, R., et al. (2013), BMP signaling controls muscle mass, *Nature Genetics*, *45*(11), 1309–1318, doi:10.1038/ng.2772.

[252] Satpathy, A. T., et al. (2019), Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion, *Nature Biotechnology*, *37*(8), 925–936, doi:10.1038/s41587-019-0206-z.

[253] Schaub, M. A., A. P. Boyle, A. Kundaje, S. Batzoglou, and M. Snyder (2012), Linking disease associations with regulatory information in the human genome, *Genome Research*, *22*(9), 1748–1759, doi:10.1101/gr.136127.111.

[254] Schep, A. N., J. D. Buenrostro, S. K. Denny, K. Schwartz, G. Sherlock, and W. J. Greenleaf (2015), Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions, *Genome Research*, *25*(11), 1757–1770, doi:10.1101/gr.192294.115.

[255] Schiaffino, S., A. C. Rossi, V. Smerdu, L. A. Leinwand, and C. Reggiani (2015), Developmental myosins: expression patterns and functional significance, *Skeletal Muscle*, *5*, 22, doi:10.1186/s13395-015-0046-6.

[256] Schmidt, F., et al. (2017), Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction, *Nucleic Acids Research*, *45*(1), 54–66, doi:10.1093/nar/gkw1061.

[257] Schulte-Merker, S., F. J. van Eeden, M. E. Halpern, C. B. Kimmel, and C. Nusslein-Volhard (1994), no tail (ntl) is the zebrafish homologue of the mouse T (Brachyury) gene., *Development*, *120*(4), 1009–1015.

[258] Schutte, N. M., I. Nederend, J. J. Hudziak, M. Bartels, and E. J. C. de Geus (2016), Twin-sibling study and meta-analysis on the heritability of maximal oxygen consumption, *Physiological Genomics*, *48*(3), 210–219, doi:10.1152/physiolgenomics.00117.2015.

[259] Scott, L. J., et al. (2016), The genetic regulatory signature of type 2 diabetes in human skeletal muscle, *Nature Communications*, *7*, 1–12.

[260] Seale, P., L. A. Sabourin, A. Girgis-Gabardo, A. Mansouri, P. Gruss, and M. A. Rudnicki (2000), Pax7 is required for the specification of myogenic satellite cells, *Cell*, *102*(6), 777–786, doi:10.1016/s0092-8674(00)00066-0.

[261] Sellick, G. S., et al. (2004), Mutations in PTF1A cause pancreatic and cerebellar agenesis., *Nature Genetics*, *36*(12), 1301–1305, doi:10.1038/ng1475.

[262] Semba, K., et al. (2013), Ectopic Expression of Ptf1a Induces Spinal Defects, Urogenital Defects, and Anorectal Malformations in Danforth's Short Tail Mice, *PLoS Genetics*, *9*(2), e1003,204, doi:10.1371/journal.pgen.1003204.

[263] Sharma, N., R. Nanta, J. Sharma, S. Gunewardena, K. P. Singh, S. Shankar, and R. K. Srivastava (2015), PI3K/AKT/mTOR and sonic hedgehog pathways cooperate together to inhibit human pancreatic cancer stem cell characteristics and tumor growth., *Oncotarget*, *6*(31), 32,039–32,060, doi:10.18632/oncotarget.5055.

[264] Sherwood, R. I., T. Hashimoto, C. W. O'Donnell, S. Lewis, A. A. Barkal, J. P. van Hoff, V. Karun, T. Jaakkola, and D. K. Gifford (2014), Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape., *Nature Biotechnology*, *32*(2), 171–178, doi:10.1038/nbt.2798.

[265] Shrikumar, A., E. Prakash, and A. Kundaje (2019), GkmExplain: fast and accurate interpretation of nonlinear gapped k-mer SVMs, *Bioinformatics*, *35*(14), i173–i182, doi:10.1093/bioinformatics/btz322.

[266] Simoneau, J. A., and C. Bouchard (1989), Human variation in skeletal muscle fiber-type proportion and enzyme activities, *The American Journal of Physiology*, *257*(4 Pt 1), E567–572, doi:10.1152/ajpendo.1989.257.4.E567.

[267] Sloan, C. A., et al. (2016), ENCODE data at the ENCODE portal, *Nucleic Acids Research*, *44*(D1), D726–732, doi:10.1093/nar/gkv1160.

[268] Spektor, R., N. D. Tippens, C. A. Mimoso, and P. D. Soloway (2019), methyl-ATAC-seq measures DNA methylation at accessible chromatin, *Genome Research*, *29*(6), 969–977, doi:10.1101/gr.245399.118.

[269] Stuart, T., et al. (2019), Comprehensive Integration of Single-Cell Data, *Cell*, *177*(7), 1888–1902.e21, doi:10.1016/j.cell.2019.05.031.

[270] Sudlow, C., et al. (2015), UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age, *PLoS Medicine*, *12*(3), doi:10.1371/journal.pmed.1001779.

[271] Sugawara, A., K. Goto, Y. Sotomaru, T. Sofuni, and T. Ito (2006), Current status of chromosomal abnormalities in mouse embryonic stem cell lines used in Japan, *Comparative Medicine*, *56*(1), 31–34.

[272] Sun, H., C. Sun, W. Xiao, and R. Sun (2019), Tissue-resident lymphocytes: from adaptive to innate immunity, *Cellular & Molecular Immunology*, *16*(3), 205–215, doi:10.1038/s41423-018-0192-y.

[273] Sun, X., and J. C. St John (2018), Modulation of mitochondrial DNA copy number in a model of glioblastoma induces changes to DNA methylation and gene expression of the nuclear genome in tumours, *Epigenetics & Chromatin*, *11*(1), 53, doi:10.1186/s13072-018-0223-z.

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

[275] Talbot, J., and L. Maves (2016), 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, doi:10.1002/wdev.230.

[276] Tang, F., et al. (2009), mRNA-Seq whole-transcriptome analysis of a single cell, *Nature Methods*, *6*(5), 377–382, doi:10.1038/nmeth.1315.

[277] Tarbell, E. D., and T. Liu (2019), HMMRATAC: a Hidden Markov ModeleR for ATAC-seq, *Nucleic Acids Research*, *47*(16), e91, doi:10.1093/nar/gkz533.

[278] Teslovich, T. M., et al. (2010), Biological, clinical and population relevance of 95 loci for blood lipids, *Nature*, *466*(7307), 707–713, doi:10.1038/nature09270.

[279] Thompson, N., E. Gésina, P. Scheinert, P. Bucher, and A. Grapin-Botton (2012), RNA profiling and chromatin immunoprecipitation-sequencing reveal that PTF1a stabilizes pancreas progenitor identity via the control of MNX1/HLXB9 and a network of other transcription factors, *Molecular and Cellular Biology*, *32*(6), 1189–1199, doi:10.1128/MCB.06318-11.

[280] Thurman, R. E., et al. (2012), The accessible chromatin landscape of the human genome, *Nature*, *488*(7414), 75–82.

[281] Thurner, M., et al. (2018), Integration of human pancreatic islet genomic data refines regulatory mechanisms at Type 2 Diabetes susceptibility loci, *eLife*, *7*, doi:10.7554/eLife.31977.

[282] Travers, M. E., et al. (2013), Insights Into the Molecular Mechanism for Type 2 Diabetes Susceptibility at the KCNQ1 Locus From Temporal Changes in Imprinting Status in Human Islets, *Diabetes*, *62*(3), 987–992, doi:10.2337/db12-0819.

[283] Uezumi, A., et al. (2014), Identification and characterization of PDGFRa+ mesenchymal progenitors in human skeletal muscle, *Cell Death & Disease*, *5*(4), e1186, doi:10.1038/cddis.2014.161.

[284] Ugarte, G., O. Cappellari, L. Perani, A. Pistocchi, and G. Cossu (2012), Noggin recruits mesoderm progenitors from the dorsal aorta to a skeletal myogenic fate, *Developmental Biology*, *365*(1), 91–100, doi:10.1016/j.ydbio.2012.02.015.

[285] Varshney, A., H. VanRenterghem, P. Orchard, A. P. Boyle, M. L. Stitzel, D. Ucar, and S. C. J. Parker (2019), Cell Specificity of Human Regulatory Annotations and Their Genetic Effects on Gene Expression, *Genetics*, *211*(2), 549–562, doi:10.1534/genetics.118.301525.

[286] Varshney, A., et al. (2017), Genetic regulatory signatures underlying islet gene expression and type 2 diabetes, *Proceedings of the National Academy of Sciences*, *114*(9), 2301–2306.

[287] Varshney, A., et al. (2020), A transcriptional regulatory atlas of human pancreatic islets reveals non-coding functional signatures at GWAS loci, *bioRxiv*, p. 812552, doi:10.1101/812552.

[288] Venables, W. N., and B. D. Ripley (2002), *Modern Applied Statistics with S*, Statistics and Computing, 4 ed., Springer-Verlag.

[289] Vlangos, C. N., B. C. O'Connor, M. J. Morley, A. S. Krause, G. A. Osawa, and C. E. Keegan (2009), Caudal regression in adrenocortical dysplasia (acd) mice is caused by telomere dysfunction with subsequent p53-dependent apoptosis., *Developmental Biology*, *334*(2), 418–428, doi:10.1016/j.ydbio.2009.07.038.

[290] Vlangos, C. N., A. N. Siuniak, D. Robinson, A. M. Chinnaiyan, R. H. Lyons, J. D. Cavalcoli, and C. E. Keegan (2013), Next-Generation Sequencing Identifies the Danforth's Short Tail Mouse Mutation as a Retrotransposon Insertion Affecting Ptf1a Expression, *PLOS Genetics*, *9*(2), e1003,205–11.

[291] Volker, L. A., et al. (2012), Comparative analysis of Neph gene expression in mouse and chicken development., *Histochemistry and Cell Biology*, *137*(3), 355–366, doi:10.1007/s00418-011-0903-2.

[292] von der Haar, T. (2008), A quantitative estimation of the global translational activity in logarithmically growing yeast cells, *BMC Systems Biology*, *2*, 87, doi:10.1186/1752-0509-2-87.

[293] Wang, H., F. Noulet, F. Edom-Vovard, F. Le Grand, and D. Duprez (2010), Bmp Signaling at the Tips of Skeletal Muscles Regulates the Number of Fetal Muscle Progenitors and Satellite Cells during Development, *Developmental Cell*, *18*(4), 643–654, doi:10.1016/j.devcel.2010.02.008.

[294] Wang, X., et al. (2017), Anti-proliferation of breast cancer cells with itraconazole: Hedgehog pathway inhibition induces apoptosis and autophagic cell death., *Cancer letters*, *385*, 128–136, doi:10.1016/j.canlet.2016.10.034.

[295] Washietl, S., M. Kellis, and M. Garber (2014), Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals., *Genome Research*, *24*(4), 616–628, doi:10.1101/gr.165035.113.

[296] Weedon, M. N., et al. (2014), Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis, *Nature Genetics*, *46*(1), 61–64, doi:10.1038/ng.2826.

[297] Wei, M., J. B. Kampert, C. E. Barlow, M. Z. Nichaman, L. W. Gibbons, R. S. Paffenbarger, and S. N. Blair (1999), Relationship between low cardiorespiratory fitness and mortality in normal-weight, overweight, and obese men, *JAMA*, *282*(16), 1547–1553, doi:10.1001/jama.282.16.1547.

[298] Wei, Z., W. Zhang, H. Fang, Y. Li, and X. Wang (2018), esATAC: an easy-to-use systematic pipeline for ATAC-seq data analysis, *Bioinformatics*, *34*(15), 2664–2665, doi:10.1093/bioinformatics/bty141.

[299] Weinstein, D. C., A. Ruiz i Altaba, W. S. Chen, P. Hoodless, V. R. Prezioso, T. M. Jessell, and J. E. J. Darnell (1994), The winged-helix transcription factor HNF-3 beta is required for notochord development in the mouse embryo., *Cell*, *78*(4), 575–588, doi:10.1016/0092-8674(94)90523-1.

[300] Welch, J. D., V. Kozareva, A. Ferreira, C. Vanderburg, C. Martin, and E. Z. Macosko (2019), Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity, *Cell*, *177*(7), 1873–1887.e17, doi:10.1016/j.cell.2019.05.006.

[301] Wiechens, N., V. Singh, T. Gkikopoulos, P. Schofield, S. Rocha, and T. Owen-Hughes (2016), The Chromatin Remodelling Enzymes SNF2H and SNF2L Position Nucleosomes adjacent to CTCF and Other Transcription Factors, *PLOS Genetics*, *12*(3), e1005,940, doi:10.1371/journal.pgen.1005940.

[302] Willet, S. G., M. A. Hale, A. Grapin-Botton, M. A. Magnuson, R. J. MacDonald, and C. V. E. Wright (2014), Dominant and context-specific control of endodermal organ allocation by Ptf1a., *Development*, *141*(22), 4385–4394, doi:10.1242/dev.114165.

[303] Willey, V. J., et al. (2018), Estimating the Real-World Cost of Diabetes Mellitus in the United States During an 8-Year Period Using 2 Cost Methodologies, *American Health & Drug Benefits*, *11*(6), 310–318.

[304] Williams, P. T. (2010), Usefulness of cardiorespiratory fitness to predict coronary heart disease risk independent of physical activity, *The American Journal of Cardiology*, *106*(2), 210–215, doi:10.1016/j.amjcard.2010.03.017.

[305] Wilson, V., P. Rashbass, and R. S. Beddington (1993), Chimeric analysis of T (Brachyury) gene function., *Development*, *117*(4), 1321–1331.

[306] Wilson, V., L. Manson, W. C. Skarnes, and R. S. Beddington (1995), The T gene is necessary for normal mesodermal morphogenetic cell movements during gastrulation., *Development*, *121*(3), 877–886.

[307] Wu, Y., et al. (2019), Colocalization of GWAS and eQTL signals at loci with multiple signals identifies additional candidate genes for body fat distribution, *Human Molecular Genetics*, *28*(24), 4161–4172, doi:10.1093/hmg/ddz263.

[308] Xi, H., et al. (2020), A Human Skeletal Muscle Atlas Identifies the Trajectories of Stem and Progenitor Cells across Development and from Human Pluripotent Stem Cells, *Cell Stem Cell*, doi:10.1016/j.stem.2020.04.017.

[309] Xian, L., et al. (2012), Matrix IGF-1 regulates bone mass by activation of mTOR in mesenchymal stem cells, *Nature Medicine*, *18*(7), 1095–1101, doi:10.1038/nm.2793.

[310] Xiong, Z., et al. (2019), Novel genetic loci affecting facial shape variation in humans, *eLife*, *8*, e49,898, doi:10.7554/eLife.49898.

[311] Yengo, L., et al. (2018), Meta-analysis of genome-wide association studies for height and body mass index in 700000 individuals of European ancestry, *Human Molecular Genetics*, *27*(20), 3641–3649, doi:10.1093/hmg/ddy271.

[312] Zanetta, L., S. G. Marcus, J. Vasile, M. Dobryansky, H. Cohen, K. Eng, P. Shamamian, and P. Mignatti (2000), Expression of Von Willebrand factor, an endothelial cell marker, is up-regulated by angiogenesis factors: a potential method for objective assessment of tumor angiogenesis, *International Journal of Cancer*, *85*(2), 281–288, doi:10.1002/(sici)1097-0215(20000115)85:2⟨281::aid-ijc21⟩3.0.co;2-3.

[313] Zaret, K. (2005), Micrococcal Nuclease Analysis of Chromatin Structure, *Current Protocols in Molecular Biology*, *69*(1), 21.1.1–21.1.17, doi:10.1002/0471142727.mb2101s69, _eprint: https://currentprotocols.onlinelibrary.wiley.com/doi/pdf/10.1002/0471142727.mb2101s69.

[314] Zaret, K. S., and J. S. Carroll (2011), Pioneer transcription factors: establishing competence for gene expression, *Genes & Development*, *25*(21), 2227–2241, doi:10.1101/gad.176826.111.

[315] Zhang, F., and J. R. Lupski (2015), Non-coding genetic variants in human disease, *Human Molecular Genetics*, *24*(R1), R102–R110, doi:10.1093/hmg/ddv259.

[316] Zhang, Y., et al. (2008), Model-based Analysis of ChIP-Seq (MACS), *Genome Biology*, *9*(9), R137, doi:10.1186/gb-2008-9-9-r137.

[317] Zhang, Y., et al. (2008), Model-based analysis of ChIP-Seq (MACS)., *Genome Biology*, *9*(9), R137, doi:10.1186/gb-2008-9-9-r137.

[318] Zhao, S., Y. Zhang, R. Gamini, B. Zhang, and D. von Schack (2018), Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion, *Scientific Reports*, *8*(1), 4781, doi:10.1038/s41598-018-23226-4.

[319] Zhao, W., X. Pan, T. Li, C. Zhang, and N. Shi (2016), Lycium barbarum Polysaccharides Protect against Trimethyltin Chloride-Induced Apoptosis via Sonic Hedgehog and PI3K/Akt Signaling Pathways in Mouse Neuro-2a Cells., *Oxidative Medicine and Cellular Longevity*, *2016*, 9826,726, doi:10.1155/2016/9826726.

[320] Zhu, C., et al. (2019), An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome, *Nature Structural & Molecular Biology*, *26*(11), 1063–1070, doi:10.1038/s41594-019-0323-x.

[321] Ziyatdinov, A., M. Vázquez-Santiago, H. Brunel, A. Martinez-Perez, H. Aschard, and J. M. Soria (2018), lme4qtl: linear mixed models with flexible covariance structure for genetic studies of related individuals, *BMC Bioinformatics*, *19*(1), 68, doi:10.1186/s12859-018-2057-x.

[322] Zuo, Z., Y. Jin, W. Zhang, Y. Lu, B. Li, and K. Qu (2019), ATAC-pipe: general analysis of genome-wide chromatin accessibility, *Briefings in Bioinformatics*, *20*(5), 1934–1943, doi:10.1093/bib/bby056.