THE UNIVERSITY OF CHICAGO

GENETIC AND GENOMIC APPROACHES TO INVESTIGATE THE ROLES OF CIS-REGULATORY ELEMENTS IN DEVELOPMENT AND DISEASE

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES

AND THE PRITZKER SCHOOL OF MEDICINE

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

COMMITTEE ON GENETICS, GENOMICS, AND SYSTEMS BIOLOGY

BY

LINDSEY ELIZABETH MONTEFIORI

CHICAGO, ILLINOIS

JUNE 2019

TABLE OF CONTENTS

List of figures viii

List of tables x

List of supplemental files (available online) xi

Acknowledgements xii

Abstract of the dissertation xiv

Chapter 1: Introduction 1

1.1 Overview of the control of gene expression regulation by cis-regulatory elements 1

1.2 Molecular characteristics of cis-regulatory elements 2

1.3 The curious case of non-coding sequence conservation 3

1.4 Gene regulation in the context of the 3D genome 5

1.5 Principles of genome organization: from large-scale compartmentalization of chromosomes to fine-scale mapping of enhancer-promoter loops 6

1.6 Gene regulation and human disease 10

1.7 Genetic variation and complex disease: The Genome-Wide Association Study (GWAS) 11

1.8 Common genetic variation and cardiovascular disease 13

1.9 Overview of thesis research projects 14

Chapter 2: Reducing mitochondrial reads in ATAC-seq using CRISPR/Cas9 17

2.1 Abstract 17

2.2 Introduction 18

2.3 Results 19

2.3.1 Development and implementation of anti-mt CRISPR treatment 19

2.3.2 Effect of removing detergent from the original ATAC-seq protocol 24

2.3.3 Variations on the anti-mt CRISPR treatment 30

2.4 Discussion 31

2.5 Methods 32

2.5.1 Human lymphoblastoid cell line growth and harvesting 32

2.5.2 Preparation of ATAC-seq libraries 33

2.5.3 Anti-mitochondrial CRISPR/Cas9 treatment 33

2.5.4 Peak calling 35

2.5.5 Fraction of TSS and enhancers intersecting peaks 36

2.5.6 Mean fraction of common peaks and mean Pearson's R^2 of read counts 36

2.5.7 Statistical tests 36

2.6 Appendix A: Supplemental Figures 38

2.7 Appendix B: Supplemental Tables 42

Chapter 3: Deletion of an ultraconserved element causes reduced body weight in mice 44

3.1 Abstract 44

3.2 Introduction 44

3.3 Results 47

3.3.1 Extreme sequence conservation and functional characterization of the Irx UCEs 47

3.3.2 Deletion of UCE5, but not UCE3, results in reduced body weight on a high fat diet .. 49
3.3.3 UCE3 and UCE5 are not required for Irx3 or Irx5 expression in the adult hypothalamus 50
3.4 Discussion 52
3.5 Methods 57
3.5.1 Generation of UCE deletion mice 57
3.5.2 High fat diet and body weight measurements 58
3.5.3 Quantitative real-time PCR 58
3.5.4 RNA-seq 59
3.6 Appendix C: Supplemental Figures 60
3.7 Appendix D: Supplemental Tables 61

Chapter 4: A promoter interaction map for cardiovascular disease genetics 69
4.1 Abstract 69
4.2 Introduction 69
4.3 Results 71
4.3.1 iPSC-derived cardiomyocytes provide an effective model to study the architecture of CVD genetics 71
4.3.2 Promoter-capture Hi-C identifies distal regulatory elements in iPSCs and CMs 72
4.3.3 Promoter interactions are enriched for tissue-specific transcription factor motifs 77
4.3.4 Long-range promoter interactions are enriched for active cis-regulatory elements and correspond to gene expression dynamics 80

4.3.5 Dynamic changes in genomic compartmentalization involve a subset of cardiac-specific genes 83
4.3.6 CM promoter interactions link GWAS SNPs to target genes 86
4.3.7 Using gene expression as a metric for interpreting disease-relevance of newly identified target genes 89
4.3.8 CM promoter interactions are informative to cardiovascular associations that do not directly involve cardiomyocytes 92
4.4 Discussion 95
4.5 Methods 100
4.5.1 Tissue culture of iPSCs 100
4.5.2 Cardiomyocyte differentiation 100
4.5.3 Promoter capture Hi-C 102
4.5.4 Interaction calling 105
4.5.5 4C-style plots 106
4.5.6 TAD analysis 106
4.5.7 A/B compartments 107
4.5.8 RNA-seq 107
4.5.9 H3K27ac ChIP-seq for comparison with Epigenome Roadmap samples 107
4.5.10 Gene Ontology analysis 108
4.5.11 Motif analysis 109
4.5.12 Histone ChIP-seq enrichment analysis 109
4.5.13 GWAS analysis 110
4.5.14 MGI analysis 111

4.5.15 eQTL analysis 111

4.5.16 Data availability 112

4.6 Appendix E: Supplemental Figures 113

4.7 Appendix F: Web links for supplemental files 122

Chapter 5: Summary and conclusions 123

Bibliography 131

LIST OF FIGURES

Figure 2.1 ATAC-seq read densities in the mitochondrial chromosome and one nuclear genome region 20

Figure 2.2 ATAC-seq was performed on human lymphoblastoid cells and half of each sample. was left untreated (green) and the other half was treated with anti-mt CRISPR 23

Figure 2.3 Effect of detergent removal from the ATAC-seq protocol 26

Figure 2.4 Comparison of the fraction of functional regions overlapping ATAC-seq peaks 28

Figure 2.5 Comparison of ATAC-seq samples normalized by total number of usable reads instead of total number of sequenced reads 29

Figure 2.6 Modifications of the anti-mt CRISPR treatment 30

Supplemental Figure S2.1 Number of peaks identified with HOMER and MACS2 using different parameters and 9.8 M, 17 M and 21.9 M reads. 38

Supplemental Figure S2.2 Fraction of peaks overlapping Epigenome Roadmap lymphoblastoid cell enhancers 39

Supplemental Figure S2.3 Effect of modifications of the anti-mt CRISPR/Cas9 treatment on the fraction of mitochondrial reads and usable reads 40

Supplemental Figure S2.4 High sensitivity Bionalyzer traces showing 3 replicates of ATAC-seq libraries before and after CRISPR/Cas9 treatment 41

Supplemental Figure S2.5 Background is higher in ND samples..... 41

Figure 3.1 Syntenic- and sequence-level conservation of the *Irx* UCEs 48

Figure 3.2 Effect of UCE3 and UCE5 deletion on body weight 49

Figure 3.3 Gene expression analysis in UCE3 and UCE5 deletion mice 51

Supplemental Figure S3.1 Generation of UCE3 and UCE5 knock-out mice 60

Figure 4.1 General features of promoter interactions 74

Figure 4.2 Transcription factor motif enrichment in distal interacting regions 79

Figure 4.3 Enrichment of promoter interactions to distal regulatory features 82

Figure 4.4 A/B compartment switching corresponds to activation of tissue-specific genes . 85

Figure 4.5 CM promoter interactions link CVD GWAS SNPs to target genes 88

Figure 4.6 Characterizing target genes based on expression level 91

Figure 4.7 Relevance of CM promoter interactions for cardiac arrhythmia,
myocardial infarction and heart failure 94

Supplemental Figure S4.1 Quality control of iPSC-CMs 113

Supplemental Figure S4.2 Analysis of RNA-seq in iPSCs and iPSC-CMs 115

Supplemental Figure S4.3 Analysis of PCHi-C interactions in the context of TADs 116

Supplemental Figure S4.4 Correlation between the number of histone
ChIP-seq peaks within 300 kb of promoters and gene expression level 118

Supplemental Figure S4.5 Comparison of A/B compartments in Hi-C and PCHi-C 119

Supplemental Figure S4.6 Example of A/B compartments 120

Supplemental Figure S4.7 GO analysis on the genes switching from active A compartments
in iPSCs to inactive B compartments in CMs 121

LIST OF TABLES

Supplemental Table S2.1 Summary of fold-differences between the number of peaks called in anti-mt CRISPR treated (TR) and untreated samples (UN), and samples prepared with (DT) and without detergent (ND) in the cell lysis buffer 42

Supplemental Table S2.2 Fold-differences between the fraction of enhancers identified in anti-mt CRISPR treated (TR) and untreated samples (UN), and samples prepared with (DT) and without detergent (ND) in the cell lysis buffer. 43

Supplemental Table S3.1 List of oligos used in the study 61

Supplemental Table S3.2 Raw high fat diet data for all animals 62

Supplemental Table S3.3 Body weight phenotypes resulting from enhancer deletions (literature search) 66

LIST OF SUPPLEMENTAL FILES (Available online)

Supplemental File S2.1 Sample information

Supplemental File S2.2 Cost estimator

Supplemental File S4.1 PCHi-C interactions for iPSC

Supplemental File S4.2 PCHI-C interactions for CM

Supplemental File S4.3 CVD SNPs

Supplemental File S4.4 HOMER motif analysis for the distal interacting regions of promoter interactions

Supplemental File S4.5 Gene Ontology enrichment output

Supplemental File S4.6 Gene Ontology input gene lists

Supplemental File S4.7 GWAS terms used to compile studies

Supplemental File S4.8 GWAS summary table

Supplemental File S4.9 Hi-C read information

Supplemental File S4.10 Public datasets used

ACKNOWLEDGMENTS

First and foremost I would like to thank my advisor, Marcelo Nobrega, for all of the guidance and support he has provided me throughout my PhD. I am particularly grateful for the opportunity to work on a project centered on genome organization, and his flexibility in allowing me to pursue and explore this interest. I really value the multiple opportunities I had to attend and present at conferences, and I have learned a lot from the guidance, input and encouragement he has given me in my projects, the writing process, and in scientific thinking in general.

I am very thankful for all of the Nobrega lab members. I couldn't have asked for a better working and scientific environment. The level of experimental expertise is outstanding and I have learned so much from you all. I would like to specifically acknowledge Debora, who really helped me get off the ground with the Hi-C work – I could not have worked on the project without out her. I also would like to thank Noboru, who patiently helped me learn the basic computational tools to be able to analyze my own data. I learned a lot under his mentorship and am thankful for his support in all of the analyses I performed. I am very grateful for my baymate, Amelia, for her wonderful friendship and great scientific discussions; these have been a bright spot in my PhD. I would also like to thank Ivy for constantly being there to discuss my experiments and for mentoring me during my rotation; Grazyna who was an amazing person to work with in the lab; Katie Bailey for showing me the ropes when I first joined the lab, and the newest graduate students, Grace and Kate – I am thankful for the opportunity to work with you two over the last year.

I would like to thank the members of my committee: Ivan Moskowitz, Yoav Gilad and Vinny Lynch, for their guidance and support. In particular, Ivan for his input on the cardiomyocyte project. I would also like to thank the professors who taught courses during our first year, especially Ilya Ruvinsky whose Evolution of Gene Regulation course will always stick with me.

I would like to thank Sue Levison, our graduate administrator, for her constant support navigating the graduate school roads, and also her friendship. I would also like to thank all of the students in my cohort, whose friendship made adjusting to graduate school life a lot easier. I am very grateful for my climbing friends who have provided an amazing outlet over the last four years: Katie Mika, Erin Fry, Charlie Lang, Soo Ji Kim and Dan Kerr. I really cherish our times together, both on and off the wall.

I would also like to acknowledge my pre-graduate experience at the NIA which set me on this path to graduate school. I am grateful for the mentorship of my previous PI, Ranjan Sen, who gave me an opportunity to learn amazing techniques while studying one of the most interesting gene regulatory phenomenon in the genome. I am also extremely grateful for the mentorship of my previous in-lab mentor, Tatiana Gerasimova, who taught me the importance of patience when it comes to experimental research, and shared her passion for science.

Finally I would like to thank my family: my mom and dad, for their constant love, support, guidance and wisdom. I feel extremely lucky to have such amazing people as parents, words cannot express how much you both mean to me and how influential you have been in my life. My sister Shannon provides a unique kind of strength that has helped me time and again, I am so thankful for having her in my life. My little sister Natalie and my step-mom Nicole have also given me immense support and love. I want to thank my best friend Kristen who has been in my life since our Raleigh days; her friendship means the world to me. I want to thank Paul, for all the fun that we have had in Chicago, being there for me through the ups and downs of graduate school and life, and for providing an unexpected twist on my scientific trajectory. Finally – I want to thank Alfie and Cosmo who make life a little warmer with their weird and unique ways.

ABSTRACT OF THE DISSERTATION

Gene regulation describes the totality of molecular events that result in precisely orchestrated gene expression patterns which collectively drive organismal development and define cellular states. Much of this logic is encoded in the genome itself, which is subject to mutation and natural variation. Because of the fundamental role that gene regulation plays in cellular biology, perturbations to gene expression patterns may have pathophysiological consequences, and these are far from being well understood. In my dissertation research, I used a variety of experimental and computational approaches to study the genetic basis of gene regulation in the context of normal cellular development and human disease. First, I developed an experimental approach to improve the ATAC-seq assay, a commonly used assay to detect regulatory elements. This approach uses CRISPR/Cas9 to remove contaminating mitochondrial DNA fragments, increasing the number of regulatory elements identified. Next, I investigated the gene-regulatory function of two ultraconserved enhancer elements in the mouse genome. I used CRISPR/Cas9 genome engineering to delete these elements from the germline and reported that deletion of one element caused a body weight phenotype, albeit in the absence of gene expression changes in the hypothalamus, challenging our view of these elements as traditional enhancers. Finally, I used promoter capture Hi-C in combination with gene expression profiling and publicly available epigenetic datasets to study the gene-regulatory changes that accompany human cardiomyocyte differentiation. I integrated these data with 50 genome-wide association study results for cardiovascular disease traits in order to prioritize target genes for functional follow up studies and provide a gene regulatory context to the thousands of loci associated with these diseases. Taken together, this work improved our ability to assay functional regions of the genome with experimental

approaches, contributed further data to the function of ultraconserved elements, and increased our understanding of the complex nature of long-range gene regulation in the context of cardiomyocyte differentiation and cardiovascular disease.

CHPATER 1: INTRODUCTION

Gene expression regulation is fundamental to all life, from the single-celled prokaryotes that turn genes on and off in response to available carbon sources, to the 37 trillion-celled humans comprised of potentially thousands of distinct cell types, each of which expresses a different pattern of genes in order to carry out its specialized function. Understanding the molecular logic governing gene expression patterns is a central tenet of gene regulation research, and a major goal is to increase our ability to interpret the functional and physiological consequences of perturbations to normal gene expression, particularly as it relates to disease. This requires the development of methods to interrogate the genome and identify regions with gene regulatory activity, functional approaches to test with certainty the *in vivo* function of putative regulatory elements such as enhancers, and general frameworks within which to consider the phenotypic impact of DNA sequence variants and other pathogenic genomic lesions on the proper functioning of the genome.

1.1 Overview of the control of gene expression regulation by cis-regulatory elements

The spatio-temporal regulation of gene expression is governed in large part by cis-regulatory elements—enhancers, repressors, and insulators—which are DNA sequences that bind transcription factors (TFs) and other molecular factors to directly influence the expression pattern of target genes. Because of their fundamental role in the control of gene expression, identifying and functionally interrogating these elements is a core goal in the field of gene regulation. Enhancers are the most widely studied class of cis-regulatory element, and are classically defined as short (< 1kb) DNA sequences that enhance or activate the expression of their target genes in response to specific developmental or environmental cues, often from great genomic distances (typically up to 1Mb)[1]. This is achieved through the combinatorial binding and subsequent delivery of tissue-specific TFs to a target gene's promoter, where direct interactions between TF activation

domains and the pre-initiation complex serve to activate or enhance transcription (reviewed in[2]). Thus, an enhancer is only "active"—meaning it actively contributes to some gene's expression—in the presence of the particular combination of TFs that bind to it. In this way, enhancers and other cis-regulatory elements serve as mediators between the cellular/extra-cellular environment and gene expression, controlling a cells' transcriptional response to varying cues in order to drive development and maintain homeostasis throughout adult life.

1.2 Molecular characteristics of cis-regulatory elements

Cis-regulatory elements may be located anywhere in the genome and are not readily identified from DNA sequence alone, which poses a challenge to locate and characterize enhancers in the various developmental/environmental contexts in which they function. Progress has been made possible by an understanding of the molecular and epigenetic characteristics of regulatory elements, and the development of tools which exploit these features to aid in their identification.

The primary biophysical characteristics of active enhancers are nucleosome depletion, which reflects TF binding to enhancer DNA, and the presence of flanking nucleosomes that contain histones with specific post-translational modifications, namely H3K27ac and H3K4me1[3–5]. In contrast, inactive enhancers and other genomic regions that have been transcriptionally silenced are often associated with nucleosomes containing H3K27me3[3,6]. The mechanistic role of these histone modifications in cis-regulatory element function is not completely understood, but likely involves the interplay between chromatin remodeling, TF recruitment, and transcriptional activation[7].

Both of these features—nucleosome depletion and post-translational histone modifications—are readily identified in genome-wide high-throughput sequencing assays.

Nucleosome-free, "open" chromatin is highly accessible to the activity of DNaseI, leading to the development of DNase-seq to map positions of accessible chromatin[8]. More recently, the Assay for Transposase-Accessible Chromatin using sequencing, or ATAC-seq, was developed to identify genomic regions tagged by the activity of the Tn5 transpose, which preferentially inserts sequencing adapters into regions of open chromatin[9,10]. ATAC-seq comes with the added benefit of requiring far fewer cells compared to DNase-seq, but suffers from technical limitations in some cell types whereby mitochondrial DNA fragments contaminate the final sequencing library. In Chapter 2 of my thesis, I present a CRISPR-based approach to remove mitochondrial reads from ATAC-seq libraries, which increases the number of enhancers identified[11]. Finally, histone modifications and TF occupancy patterns are assessed using chromatin immunoprecipitation followed by sequencing (ChIP-seq), which relies on antibody-mediated pull-down of DNA sequences associated with or bound by these molecular factors.

As a testament to the importance of this suite of technologies in mapping gene-regulatory landscapes, DNase-seq, ATAC-seq, and ChIP-seq have now been conducted in hundreds of different cell types across several species, representing diverse developmental stages and disease states[12–15]. These datasets collectively describe a highly complex and dynamic genomic environment consisting of an estimated 400,000 enhancers[7]. A daunting task facing the field is to understand the function of each of these enhancers: When is a given enhancer active? Which gene does it regulate? What determines which gene it regulates? Is the enhancer required for target gene expression? What are the functional consequences of mutations within enhancers?

1.3 The curious case of non-coding sequence conservation

In addition to the biophysical characteristics described above, nucleotide sequence conservation has been used to identify putative cis-regulatory elements under the assumption that

natural selection would retain the TF binding sites required for their function[16,17]. However, studies performed in diverse species from fly to human have shown quite the opposite: important developmental and tissue-specific enhancers are often not conserved at the sequence level despite functional conservation at the gene regulatory level[18]. For example, the functionally conserved *even-skipped* (*eve*) enhancers in *Drosophila melanogaster* and five other scavenger fly species are not conserved at the sequence level, despite driving the same pattern of *eve* expression in the developing embryo[19]. Furthermore, Blow *et al.* identified many putative heart enhancers in mice that are only weakly conserved despite showing robust evidence of heart enhancer activity at a developmentally constrained time-point[20]. One interpretation resulting from these studies is that the overall combination of TF binding sites within an enhancer remains stable despite changes to the one-dimensional nucleotide sequence, ensuring conserved enhancer activity in the face of sequence mutation[21]. Nevertheless, there exist thousands of highly conserved non-coding elements in the human genome which are hypothesized to act as critical developmental and tissue-specific enhancers[17,22–24]. One challenge is to reconcile our knowledge regarding enhancer sequence turnover with the nucleotide sequence conservation imposed by functional constraint.

A class of non-coding elements termed ultraconserved elements (UCEs) epitomizes this dilemma and will be the subject of Chapter 3 of this thesis. UCEs were originally identified in a 2004 study that used a human-mouse-rat sequence alignment to search for genomic regions of at least 200 consecutive nucleotides with 100% sequence identity[22]. These criteria identified 481 UCEs in the human genome, most of which are non-coding. Additionally, many UCEs are located near developmentally important genes such as transcription factor genes, and are often in syntenic regions in diverse species[25] which strongly suggests that they play critical roles in regulating the expression of these genes. In support of this, over half of UCEs tested in *in vivo* mouse transgenic

reporter assays displayed tissue specific activity[24]. However, in spite of the abundant functional data, germline deletion of 5 out of 8 UCEs tested in mice to date have failed to cause a phenotype[26–28]. This raises two important points: (i) sequence conservation is not a sufficient metric to gauge whether or not an element is a stereotypical enhancer, as supported by the fact that many enhancers do not show sequence conservation, and (ii) genomic deletion of a presumed enhancer is necessary to assess its true *in vivo* role as an enhancer. Although we have a wealth of correlative epigenetic data that identifies hundreds of thousands of putative enhancers, it is not possible to say with certainty that any give DNA sequence functions as an enhancer until it is deleted or otherwise disrupted in a genomic context. In Chapter 3, I use CRISPR genome editing to delete two UCEs from the mouse genome in order to assess their functional importance in the regulation of their presumed target genes, *Irx3* and *Irx5*.

1.4 Gene regulation in the context of the 3D genome

The epigenetic landscape describes the genome in a one-dimensional space, however gene expression occurs within the context of the broader nuclear environment, where chromatin is highly folded in order to fit the 2 meters of DNA into a roughly 10 μm^2 nuclear volume. One critical consequence of this organization is that distal regions of the genome are brought into physical proximity, enabling communication between distal regulatory elements and target gene promoters. In order to grasp the logic underlying enhancer function, it is first necessary to consider gene regulation in the context of this spatial genomic organization.

Clues that genome organization played a role in gene regulation originated from early studies using fluorescent in situ hybridization (FISH) to localize specific genomic regions in single nuclei. For example, Kosak *et al*. showed that the immunoglobulin heavy chain (*IgH*) locus is relocated from the nuclear periphery—typically a transcriptionally repressive environment—to the

nuclear interior consequent with the onset of V(D)J recombination in developing B cells[29]. Subsequently, it was shown that this repositioning coincides with the large-scale contraction of the ~3 Mb *IgH* locus to facilitate recombination of the V, D and J gene segments[30]. A correspondence between nuclear location and transcriptional activity has been noted for other loci as well, including the *Mash1* gene which is repositioned during neuronal development[31], and the *c-maf* gene which is sequestered to the nuclear periphery during T helper cell commitment[32]. These studies helped lend support to the emerging notion that genome organization and spatial nuclear context are important factors in the regulation of gene expression.

Despite the power of FISH technology to analyze locus dynamics in single cells, it is a relatively low-throughput technique and unable to survey genome organization at higher levels. In 2002, the chromosome conformation capture (3C) technique was developed which enabled detection of chromatin contacts at unprecedented resolution (on the order of 1-8 kb)[33]. Within seven years, this technique had been expanded upon to include the rest of the "C" technologies: 4C, 5C, and Hi-C[34–36]. Each iteration of 3C enabled a more systematic survey of chromatin contacts throughout the genome and quickly led to a doctrine whereby long-range regulation of gene expression was a pervasive phenomenon in many organisms[1,37–39]. When considered together with epigenetic maps of chromatin accessibility and histone modifications, it becomes possible to conceptualize how hundreds of thousands of cis-regulatory elements are able to control expression of distal target genes in a highly precise, dynamic manner.

1.5 Principles of genome organization: from large-scale compartmentalization of chromosomes to fine-scale mapping of enhancer-promoter loops

We now understand that the 3D genome is a major player in gene regulation, and likely contributes to gene expression in ways we still do not full grasp; however, several key principles

of genome organization have been worked out in the past 5-10 years that serve as a framework within which to investigate gene regulation by distal enhancer elements. The first key finding was enabled by the development of Hi-C, the 3C-based method that detects long-range chromatin contacts on a genome-wide scale. Hi-C experiments performed in a human lymphoblastoid cell line revealed that the genome is organized into two "compartments", termed "A" and "B", which correspond to open and closed chromatin, respectively[36]. These compartments were identified on the basis of the correlation of interaction frequencies across and between entire chromosomes, indicating that large swaths of the genome preferentially associate with either the active or inactive compartment. This demarcation of the genome based on interaction frequency is strikingly similar to microscopy-based observations which divide the genome into euchromatin and heterochromatin on the basis of chromatin density. Not unsurprisingly, the A compartment corresponds to euchromatin, whereas the B compartment corresponds to heterochromatin; moreover, genomic regions dynamically switch compartments during cell differentiation in accordance with activation/repression of genes driving differentiation[40].

The next major insight into genome organization was borne out of both Hi-C and 5C technologies applied to embryonic stem cells. Analysis of Hi-C data at a relatively high resolution of 40 kb revealed that chromosomes are organized into series of "topologically associated domains", or TADs, which are defined as genomic regions where the interaction frequencies within the region are greater than between adjacent regions[41]. When viewed as a two-dimensional interaction matrix, TADs are easily discerned as characteristic ~1 Mb triangles along the diagonal. In a separate study, Nora *et al.* used the 5C technique to survey the organization of the active and inactive X chromosomes in mouse embryonic stem cells and found the same organizational principle[42].

A significant functional consequence of a TAD-based chromosome structure is that the action of cis-regulatory elements is likely to be constrained to genes within the same TAD. For example, there are on average 12 genes per TAD in the human genome, vastly limiting the search space for an enhancer to find its target gene. In support of this notion, TADs were shown to correspond to gene-regulatory blocks which are genomic regions containing a set of enhancers critical for some target gene's expression pattern[43]. These blocks tend to remain intact throughout evolution, and many are present in syntenic regions of greatly diverged species. Furthermore, computational prediction of enhancer-target gene pairs using ATAC-seq data showed that a majority (74%) were located within the same TAD[44], again suggesting that TADs encompass most enhancer-gene interactions. Whether TAD structures evolved to maintain gene-regulatory blocks, or as a consequence of ancestral enhancer landscapes is not known, however it is clear that TADs are important for maintaining proper gene expression regulation as disruption of TAD boundaries has been reported in several studies to cause aberrant expression patterns which may underlie disease[42,45–48].

A final critical insight into the relationship between genome organization and gene expression emerged from the development of the promoter capture Hi-C (PCHi-C) technique, a sequence capture-based method to enrich Hi-C sequencing libraries for fragments mapping to promoter regions[49,50]. PCHi-C enables analysis of all promoter interactions in a population of cells at enhancer-level resolution which revealed that—in contrast to TADs—promoter interactions are highly dynamic and reflect cell differentiation state and gene expression status, although the majority of these interactions do occur within the boundaries of topological domains[51,52].

Several PCHi-C studies reported that promoter interactions are enriched for putative enhancers, with a positive relationship between gene expression and long-range interactions to

genomic regions marked by H3K27ac[49,52]. Furthermore, many enhancer interactions were cell-type specific, reflecting dynamic gene expression patterns. These observations paint a picture whereby long-range interactions between an enhancer and its target promoter are regulated such that enhancer-promoter interactions only occur when the enhancer actively contributes to gene expression. Although this observation, and the resulting model, fits the prevailing hypothesis surrounding enhancer function, studies in fly, mouse, and human have shown that many enhancer-promoter interactions are stable, in that they do not dynamically form at the onset of gene expression[53–55]. Regardless of the logic underlying the formation of enhancer-promoter interactions, most experimental evidence supports that enhancers activate their target genes via looping, indicating that physical proximity is necessary for their effect[56–59]. For example, recruitment of a synthetic transcription factor to the promoter of the adult β-globin gene resulted in formation of a long-range loop between the promoter and its distal enhancer and coincided with reactivation of normally silent β-globin gene expression[60], suggesting that forcing an interaction between a promoter and its enhancer is sufficient to activate gene expression. Moreover, this phenomenon was directly visualized in *Drosophila* embryos using live-cell imaging techniques to measure the timing of an enhancer-promoter interaction with gene transcription—the key result being a total dependence on sustained enhancer-promoter interaction to produce transcripts[61].

It is relatively straight forward to obtain correlative evidence from PCHi-C that gene expression is influenced by the action of distal-acting enhancers, yet much more technically challenging to test the necessity of an enhancer for target gene expression. However, the experimental evidence so far strongly supports that gene expression is directly influenced by enhancers through long-range DNA looping interactions[56–59]. As such, we may assume that techniques such as PCHi-C are able to identify putative enhancers and connect them to their target

gene(s) by virtue of the evidence of a physical interaction between these elements. The ability to analyze, in an unbiased and high-resolution manner, all genomic regions contacting all gene promoters is a powerful tool to integrate gene expression, epigenetic maps and genome organization.

1.6 Gene regulation and human disease

Due to the fundamental role that gene regulation plays in cellular development and tissue homeostasis, perturbations to gene regulation—either through alterations in spatio-temporal expression patterns, or through changes in the levels at which genes are supposed to be expressed—can cause pathologic phenotypes. Genomic lesions that affect cis-regulatory elements are a prime suspect in these cases. For example, point mutations in a 1 Mb-distal enhancer of the *SHH* gene are sufficient to alter *SHH* expression and cause polydactyly in humans[62]. Similarly, a single nucleotide mutation in an enhancer for the *TBX5* gene was found to alter *TBX5* expression during heart development, leading to congenital heart defects[63]. Developmental disorders such as these are relatively rare, and pinpointing the causal mutations is feasible with classic linkage analysis followed by deep targeted sequencing. Moreover, congenital developmental disorders are typically caused by alterations to a single gene, which clarifies the genetic basis of disease etiology. On the other hand, the genetic contribution to the exceedingly common complex diseases such as type 2 diabetes, obesity, autoimmune disease, and cardiovascular disease is thought to derive from small changes to the expression regulation of many genes[64,65].

The genetic contribution to complex disease is rooted in natural genetic variation, i.e. single nucleotide polymorphisms (SNPs) and structural variants (copy number alterations or small inversions and deletions) that segregate in human populations. There are over 100 million SNPs in the present global population, with each individual carrying between 4 and 5 million SNPs in

their genome[66]. Not unsurprisingly, the majority of these SNPs are non-coding, as variation within protein coding genes is likely to be less tolerated and the percentage of the genome that codes for protein is vastly smaller than the non-coding percentage (2% vs. 98%). As outlined above, non-coding cis-regulatory elements function by binding sequence-specific transcription factors in order to regulate the expression of their target genes. Thus, sequence-level alterations to these elements, including SNPs, are likely to impact their activity and, consequently, alter target gene expression. Indeed, the genotype and tissue expression (GTEx) consortium has identified several hundred thousand expression quantitative trait loci (eQTLs), which are SNPs and other DNA variants that associate with variation in gene expression levels; in total, there is at least one eQTL for nearly every gene in the human genome[67]. This principle whereby genetic variation impacts regulatory element function and target gene expression has become a tenet of modern-day genomics research which seeks to understand the phenotypic consequences of human genetic variation in both health and disease.

1.7 Genetic variation and complex disease: The Genome-Wide Association Study (GWAS)

In order to identify genetic variants that may functionally contribute to the development of complex diseases, researchers perform a genome-wide association study, or GWAS. The study design includes tens to hundreds of thousands of individuals split into either the cases (affected) or control group. Dense genotyping data from each individual is used to statistically test ~0.5-2 million SNPs for independent association with either group. In this way, a GWAS identifies regions of the genome where a SNP is statistically more likely to be found among individuals with the disease compared to those without the disease, with the implication being that the SNP functionally contributes to the disease. The power of this approach to identify variants associated

with complex disease is indicated by the now thousands of SNPs associated with hundreds of different human diseases[68].

A key challenge facing the field is to understand how to interpret disease-associated variants, ideally in a way that can be used diagnostically or even therapeutically. Perhaps the most pressing issue is to identify the gene or genes whose expression is likely influenced by each of the individual SNPs associated with a disease. For example, there are over 700 SNPs associated with various cardiovascular diseases (CVDs), and greater than 90% of these variants are in the non-coding genome and are located far from protein-coding genes. Thus, the same issues facing identification of target genes for cis-regulatory elements also apply to disease-associated variants. Indeed, across all diseases and complex traits analyzed to date, the overwhelming majority of associated SNPs are non-coding, and are frequently located in regions of open chromatin[69] and correspond to eQTLs[70], reflecting that it is gene regulation—not protein alterations—which forms the genetic basis of complex disease. By identifying the true target gene(s) for disease-associated variants, it will be possible to begin to formalize hypotheses regarding the molecular basis of complex disease.

A preeminent example which highlights this challenge is the case of *FTO*. This gene harbors the strongest genetic association with obesity risk, which is located within a ~40 kb region of the first intron[71]. Because the associations mapped within *FTO*, researchers focused heavily on investigating the role of *FTO* in obesity biology[72,73]. However, by using a combination of gene expression and chromosome conformation capture approaches, Smemo *et al*. demonstrated that the variants within the associated region alter the expression of a much more distal gene, *IRX3*, and not *FTO*[74]. Furthermore, they showed in mouse models that *Irx3* contributes to body weight homeostasis through its action in the hypothalamus. This novel insight into obesity biology would

not have been realized without first identifying a causal genomic region through GWAS and subsequent identification of the true target gene through integrated genomics approaches, including 3C-based technology.

The study by Smemo *et al*. helped solidify the importance of analyzing gene regulation in the context of the 3D genome, sending a note of caution to the common practice of assuming a disease-associated SNP is functionally connected to the closest or host gene. Indeed, multiple studies have recently used chromatin conformation data to link GWAS-identified SNPs to their likely target genes[52,75–80]. For example, Javierre *et al.* generated PCHi-C in 17 different immune cell populations and used these data to link GWAS variants associated with immune disorders to over 2,000 putative target genes[52]. Importantly, the authors identified target genes with well-known roles in immune disease pathology, as well as many genes with as-yet-unknown roles in immune-related disorders. Similar to Smemo *et al*, this study highlights the potential for identifying genes with novel roles in disease pathology, an important first step to help interpret the diagnostic and therapeutic benefit of identifying and characterizing GWAS variants.

1.8 Common genetic variation and cardiovascular disease

As mentioned previously, there are hundreds of genomic loci associated with increased risk for numerous cardiovascular diseases, including arrhythmias, heart failure, and myocardial infarction (https://www.ebi.ac.uk/gwas/). Collectively, CVDs kill more people world-wide each year compared to any other disease or cause of death. A large contributor to this wide-spread disease is driven by adoption of the so-called Western diet and lifestyle, which increases the risk for CVD through poor health choices. Despite the known influence of environmental factors on disease risk, GWASs have consistently identified strong genetic associations with CVDs across

populations, including when individuals are classified according to their lifestyle habits[81], underscoring the need to investigate the genetic component of these diseases.

One of the first studies to mechanistically dissect a CVD GWAS locus was in 2010, where Musunuru *et al.* investigated the strongest association for myocardial infarction[82]. The association was in the short, intergenic region between two genes, *CELSR2* and *PSRC1*. The authors used eQTL analysis and enhancer reporter assays to determine that the associated interval acted as a regulatory element for the more distal *SORT1* gene; moreover, a SNP identified in the myocardial infarction GWAS altered the activity of this enhancer which affected *SORT1* expression levels. The authors went on to show that *SORT1* directly controls circulating LDL cholesterol levels, providing a mechanistic link between the non-coding GWAS association and disease risk. The impact of this study derived from the authors' meticulous use of functional assays to uncover the true target gene, and highlights the dire need for systematic approaches to comprehensively map the full catalog of CVD GWAS associations to their functional targets. In Chapter 4 of my thesis, I integrate many of the aforementioned aspects of gene regulation, including epigenetic data, gene expression, and genome organization, in order to functionally connect hundreds of CVD-associated variants to putative target genes in human cardiomyocytes[83].

1.9 Overview of thesis research projects

The bulk of my thesis research has focused on addressing outstanding problems in the field of gene regulation and can be divided into three chapters related to methods development (Chapter 2), enhancer function (Chapter 3), and disease genetics (Chapter 4).

Chapter 2: Most key insights into the inner workings of gene regulation, both on a fundamental cell development level as well as in the context of disease, have been the direct result

of new technologies that enable us to study gene regulation at increasing resolutions and throughput. One such technology is ATAC-seq, the genome-wide assay to map open chromatin in any cell of interest. As mentioned in Chapter 1.2, a key limitation of ATAC-seq is contamination with mitochondrial DNA fragments which are incorporated into the final sequencing library. It has been shown that unwanted fragments can be removed from a sequencing pool by designing guide RNAs against these fragments and treating with the Cas9 endonuclease[84]. In Chapter 2, I describe our approach to remove mitochondrial DNA fragments from ATAC-seq libraries[11]. This treatment reduces the overall cost associated with sequencing, and increases identification of enhancer elements in lymphoblastoid cells.

Chapter 3: Large-scale efforts from ENCODE, the Epigenome Roadmap Project, FANTOM, and others have comprehensively mapped the positions of putative enhancers in hundreds of different cell types. These data give context to the regulation of gene expression across tissues and developmental stages and also aid the interpretation of disease-associated genetic lesions. However, these data are entirely descriptive and do not inform on the *in vivo* activity of any given putative enhancer, nor do they reveal which enhancers actively contribute to gene expression. It is becoming clear that in order to address these issues, deletion studies are required to functionally test the requirement of a predicted enhancer for gene expression regulation[85]. This is exemplified by the observation that deletion of several of the most conserved enhancer elements in the mouse genome does not result in overt gene expression or developmental phenotypes, despite displaying all of the hallmarks of critical enhancers[26]. In Chapter 3, I use CRISPR/Cas9 technology to delete two ultraconserved elements from the mouse genome (Montefiori *et al*, in preparation). I show that neither deletion affects hypothalamic expression of the predicted target genes, *Irx3* and *Irx5*, despite the observation that deletion of one of the elements causes a body

weight phenotype. These data support that *in vivo* enhancer deletions are necessary to ascertain the functional relevance of an enhancer for gene expression regulation in a particular context.

Chapter 4: One of the most pressing issues in human genetics research is to develop a better understanding of the genetic basis of complex disease. Achieving this goal requires knowledge about the gene or genes whose dysregulation drives disease etiology; however, this is complicated by the fact that the vast majority of GWAS loci are located in the non-coding genome and their target genes are not known. In Chapter 4, I use the PCHi-C technique in combination with gene expression and epigenetic data to analyze CVD GWAS loci and link them to their most likely target gene(s)[83]. I make use of an established *in vitro* cardiomyocyte differentiation protocol to ensure that the gene regulatory landscape I analyzed is representative of human cardiomyocytes. Through extensive quality control and comparative analyses, I showed that long-range promoter interactions are enriched for genomic regions that display cell-type-specific enhancer marks, and I interpret the function of these contacts on gene expression regulation. Analysis of PCHi-C interactions in cardiomyocytes identified target genes for 1,999 CVD-associated SNPs. These target genes include known CVD modifiers as well as genes with potentially novel roles in CVD biology. Importantly, all data generated in this study are freely available and easily accessible in a public genome browser, enabling easy access to the cardiovascular research community.

CHAPTER 2: REDUCING MITOCHONDRIAL READS IN ATAC-seq USING CRISPR/Cas9

2.1 Abstract[1]

ATAC-seq is a high-throughput sequencing technique that identifies open chromatin. Depending on the cell type, ATAC-seq samples may contain ~20-80% of mitochondrial sequencing reads. As the regions of open chromatin of interest are usually located in the nuclear genome, mitochondrial reads are typically discarded from the analysis. We tested two approaches to decrease wasted sequencing in ATAC-seq libraries generated from lymphoblastoid cell lines: targeted cleavage of mitochondrial DNA fragments using CRISPR technology and removal of detergent from the cell lysis buffer. We analyzed the effects of these treatments on the number of usable (unique, non-mitochondrial) reads and the number and quality of peaks called, including peaks identified in enhancers and transcription start sites. Both treatments resulted in considerable reduction of mitochondrial reads (1.7 and 3-fold, respectively). The removal of detergent, however, resulted in increased background and fewer peaks. The highest number of peaks and highest quality data was obtained by preparing samples with the original ATAC-seq protocol (using detergent) and treating them with CRISPR. This strategy reduced the amount of sequencing required to call a high number of peaks, which could lead to cost reduction when performing ATAC-seq on large numbers of samples and in cell types that contain a large amount of mitochondria.

[1] Reproduced with permission from: Montefiori, L. *et al.* Reducing mitochondrial reads in ATAC-seq using CRISPR/Cas9. *Sci. Rep.* **7,** 2451 (2017)

2.2 Introduction

ATAC-seq aims at identifying DNA sequences located in open chromatin, i.e., genomic regions whose chromatin in not densely packaged and that can be more easily accessed by proteins than closed chromatin. The ATAC-seq technique makes use of the Tn5 transposase, an optimized hyperactive transposase that fragments and tags the genome with sequencing adapters in regions of open chromatin[9]. The output of the experiment is millions of DNA fragments that can be sequenced and mapped to the genome of origin for identification of regions where sequencing reads concentrate and form "peaks".

While ATAC-seq often generates high-quality data with low background, certain cell types and tissues yield an enormous fraction (typically 20-80%) of unusable sequences of mitochondrial origin. In order to reduce the amount of wasted sequencing reads, targeted cleavage of DNA fragments has recently been used to deplete mitochondrial ribosomal RNA-derived fragments in RNA-sequencing libraries[86]. In another study, Wu *et al.* targeted the mitochondrial genome in ATAC-seq experiments using 114 guide RNAs (gRNAs) and observed a ~50% decrease in mitochondrial reads and no adverse modification of the read enrichment pattern[87].

To analyze the effect of this approach on the quality of the data, we designed 100 gRNAs targeting the human mitochondrial chromosome every ~250 base pairs (bp) and treated lymphoblastoid cell line ATAC-seq sequencing libraries with these gRNAs and Cas9 enzyme[88], hereafter referred to as anti-mt CRISPR. We compared this method to a modified ATAC-seq protocol that also aims at reducing mitochondrial reads by removing detergent from the cell lysis step, which is believed to prevent lysis of the mitochondrial membrane[89].

We observed that while both methods considerably reduced the number of mitochondrial reads sequenced, each method displayed different effects on the number of peaks called. Whereas

the removal of detergent from the lysis buffer had the largest effect in reducing mitochondrial reads, it resulted in decreased quality of the ATAC-seq libraries, as measured by the number of peaks called at a given sequencing depth, the total number of reads in peaks, and the fraction of transcription start sites (TSSs) and enhancers identified. Conversely, in addition to decreasing the number of mitochondrial reads, the anti-mt CRISPR treatment also resulted in a greater number of peaks, a greater number of reads in peaks, and higher overlap of peaks with TSSs and enhancers. Performing anti-mt CRISPR requires the one-time purchase of gRNA template oligos, as well as purchase of the Cas9 enzyme. However, the gRNAs can be generated from template DNA oligos indefinitely and shared as a community resource, potentially trivializing the up-front cost. Laboratories generating large numbers of ATAC-seq experiments on cell types that yield a high fraction of mitochondrial reads could benefit from mitochondrial depletion to decrease the cost of sequencing.

2.3 Results

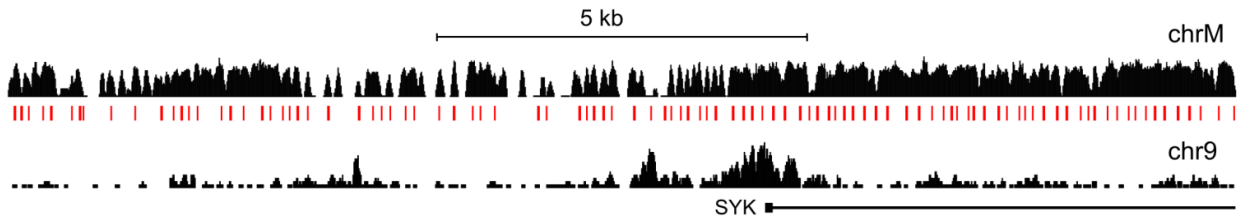
2.3.1 Development and implementation of anti-mt CRISPR treatment

The anti-mt CRISPR treatment consisted of 100 guide RNAs (gRNAs) targeting the human mitochondrial genome at regular intervals, which is usually densely covered by ATAC-seq reads generated from lymphoblastoid cell lines (LCLs), as shown in Figure 2.1. The rationale was to cleave targeted DNA fragments in the sequencing library, rendering them unable to bind and amplify on the Illumina HiSeq flow cell. Similarly to Gu *et al*.[86] and Wu *et al.*[87], we chose to treat the final (PCR amplified) sequencing library with the gRNA/Cas9 mix instead of the unamplified tagmented DNA because of the small amount of DNA present in the sample at this earlier step. Although treating the samples before PCR amplification might result in lower fractions of

mitochondria, we chose the conservative approach of treating larger amounts of DNA to reduce technical variability.

To develop and analyze the anti-mt CRISPR treatment, we used 50,000 human LCLs per sample and generated a total of 27 pairs of ATAC-seq libraries for Illumina high-throughput sequencing according to the protocol of Buenrostro *et al.*[10] (Array Express accession number E-MTAB-5205 and Supplemental File S2.1). We split each of the 27 libraries into two equal parts, leaving one half untreated and treating the other half with 100 mitochondrial gRNAs and Cas9. Due to the single turn-over nature of Cas9, Gu *et al.*[86] used an excess of enzyme and of gRNA to deplete mitochondrial ribosomal DNA. Based on this notion, we used 100X Cas9 and 100X gRNA excess. We assumed 50% mtDNA fragments in the PCR-amplified ATAC-seq library to calculate exact amounts to be used in the treatment (see section 2.5 Methods).

Figure 2.1 ATAC-seq read densities in the mitochondrial chromosome and one nuclear genome region.
Top: The mitochondrial chromosome (chrM) is densely covered by uniquely mapped reads. Genomic location of the 100 mitochondrial guide RNAs (red tick marks) designed to target the human mitochondrial chromosome (top). Bottom: compare chrM to a 16.5 kb region of the nuclear genome (hg38, chr9:90,791,567-90,808,137). The chrM and chr9 tracks are shown in different height scales for easier visualization. Seven samples were pooled and 227 M reads were sampled.



We obtained between 9.8 M and 108.6 M reads per sample in four batches of experiments. Because different numbers of reads were sequenced from each sample due to imprecision in DNA quantification and the number of multiplexed samples, we randomly sampled a fixed number of

sequenced reads from each library in order to compare across samples. This approach allowed us to assess which library preparation method yielded the best results regardless of how it affected the number of aligned or usable (unique, non-mitochondrial) reads.

After aligning reads to the human genome, we removed mitochondrial reads and reads aligned to identical coordinates and called peaks using HOMER[90] and MACS2[91]. Qualitatively similar results were obtained with both peak callers at three read depths (9.8 M (54 samples), 17 M (52 samples) and 21.9 M (47 samples)) and using different parameters to call peaks (Supplemental Figures S2.1 and S2.2). The results reported in the figures were obtained with MACS2 using custom parameters and 21.9 M sequenced reads. Results for all other read depths and parameters are presented in Supplemental Figures S2.1 and S2.2 and Supplemental Tables S2.1 and S2.2.

Figure 2.2 shows the comparison between 14 ATAC-seq samples before and after treatment with anti-mt CRISPR, using the original ATAC-seq protocol that includes detergent (DT). Visual inspection of the data showed that the untreated and treated samples were similar (Figure 2.2A), indicating that the treatment did not damage the samples. As expected, the anti-mt CRISPR treatment resulted in depletion of mitochondrial reads, while the number of reads in the nuclear genome increased (Figure 2.2B).

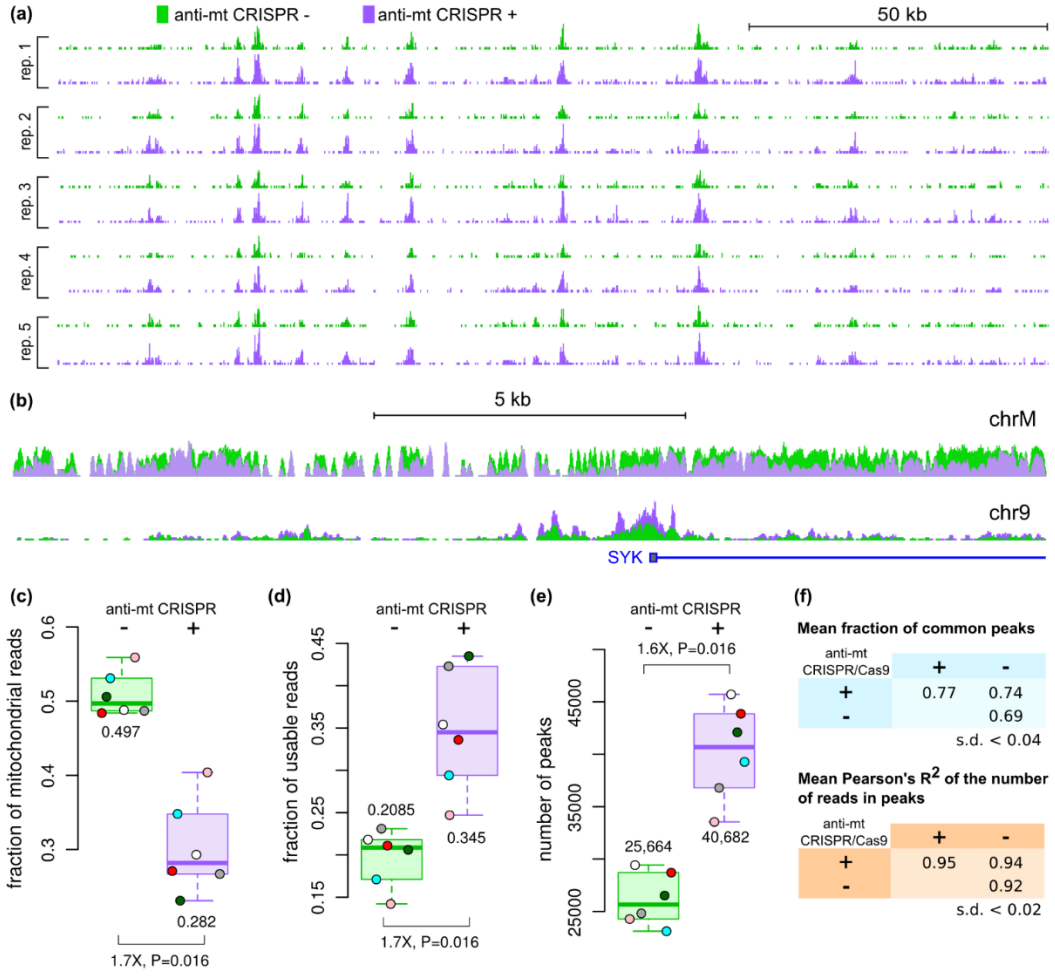
At the same sequencing depth, the anti-mt CRISPR-treated samples yielded considerably less mitochondrial reads (Figure 2.2C). This result is similar to the level of reduction of ~50% reported by Wu *et al.*[87]. Consequently, more usable reads (non-mitochondrial reads with unique coordinates), were generated (Figure 2.2D and Supplementary Figure S2.1). The increased number of usable reads resulted in 50% more peaks in the treated halves of all samples (Figure 2.2E and

Supplementary Figure S2.2), demonstrating the importance of removing excess mitochondrial reads from ATAC-seq samples.

One concern when treating samples with CRISPR/Cas9 was whether off-target gRNA/Cas9 activity would affect the data to a significant extent. To address this issue, we compared the percentage of peaks common across replicates (1 bp overlap) and across anti-mt CRISPR-treated and untreated samples. Figure 2.2F shows that the degree of overlap between untreated and treated samples was not smaller than the degree of overlap between replicates of the same condition, indicating that the anti-mt CRISPR treatment did not cause loss of peaks or create artefactual peaks. This observation is in accordance with a previous report that CRISPR treatment of sequencing libraries did not modify the read enrichment pattern[87]. We also found evidence that the anti-mt CRISPR-treated samples identified more transcription start sites and enhancers than untreated samples (see below), indicating that mtDNA cleavage did not negatively affect the data. Analysis of samples normalized by the number of usable reads instead of total number of reads sequenced (see below), corroborates the idea that the anti-mt CRISPR does not damage ATAC-seq samples.

Figure 2.2 ATAC-seq was performed on human lymphoblastoid cells and half of each sample was left untreated (green) and the other half was treated with anti-mt CRISPR (purple).

(A) Representative genomic region (hg38, chr2:74,425,417-74,586,546) showing read counts (usable reads) in 5 replicate pairs (DT) at the same sequencing depth of 21.9 M reads. Differences between treated and untreated samples were minimal, indicating that the treatment did not damage the samples. **(B)** ATAC-seq reads in the mitochondrial chromosome and in a 16.5 kb region of chromosome 9 around the SYF promoter (same as Figure 2.1). For each condition, all samples were pooled together and 227 M reads were sampled. **(C)** Treated samples yielded 1.7-fold fewer mitochondrial reads compared to untreated samples. **(D)** Accordingly, the number of unique, non-mitochondrial (usable) reads was 1.7-fold higher in treated samples than in their untreated counterparts. **(E)** At the same sequencing depth, 1.6-fold more peaks were called in the treated samples. Only 6 data points are shown because the treated halves of samples 18 and 19 (same batch) had only 14.5 M and 9.8 M reads each and were combined for improved peak calling. **(F)** Anti-mt CRISPR-treated samples shared a similar number of peaks with treated replicates and untreated samples. The top 20,000 peaks of each sample were used in this analysis. Comparison of peaks at the read count level also supports that peaks from treated samples do not substantially differ from untreated samples. Fold-differences were calculated on the medians. **(C-E)**: all samples normalized to 21.9 M sequenced reads.



2.3.2 Effect of removing detergent from the original ATAC-seq protocol

Another method that has been used to reduce the fraction of mitochondrial reads is the removal of detergent from the cell lysis step of the ATAC-seq protocol[89]. We generated seven ATAC-seq samples with no detergent (ND) and observed several differences compared to the original protocol with detergent (DT).

Interestingly, the fraction of unique reads was considerably higher in ND samples compared to DT samples (56.5% vs. 32.5%, respectively), which could reflect a lower fraction of mitochondrial fragments before PCR amplification of the sequencing library. In addition, the fraction of reads uniquely aligned to the genome was slightly higher in ND samples, compared to samples prepared with the original protocol (83.6% vs. 74.6%, respectively). This difference is due to discarding reads that map to both mitochondrial and nuclear genomes (6% of ND reads and 17% of DT reads) in order to retain only uniquely aligned reads. Because we started our analyses with the same number of sequenced reads, these differences in mappability were accounted for in our comparisons.

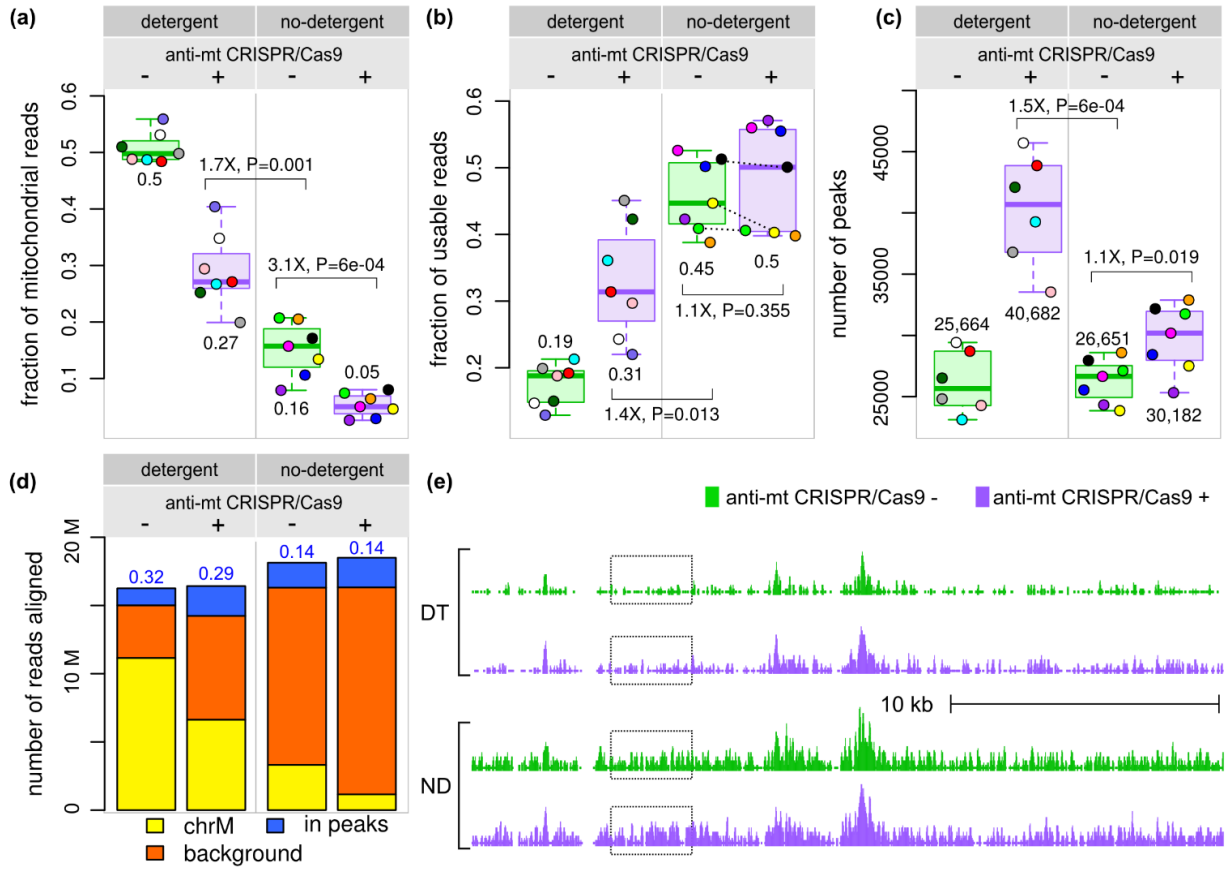
Figure 2.3 shows that the removal of detergent had a pronounced depletive effect on mitochondrial reads compared to untreated DT libraries (Figure 2.3A) and consequently increased the fraction of usable reads (Figure 2.3B). Despite this 2.4-fold increase of ND usable sequences (0.45/0.19), the number of peaks called was higher by only 1.04-fold compared to untreated DT samples (Figure 2.3C, 26,651/25,664). The mean fold-difference using other parameters to call peaks and read depths was higher at 1.2-fold, but still substantially lower than the increase in usable reads (Supplemental Table S2.1). This difference could be due to the increased background in ND samples (Figures 2.3D and 2.3E), as suggested previously[89]. We considered background reads as reads that were not mitochondrial and were not in any ATAC-seq peak identified in any

DT or ND sample. The lower signal/noise ratio in ND samples (Figure 2.3D) provides an explanation for why fewer peaks were identified in ND samples. Thus, although removing detergent from the lysis buffer increased the overall number of non-mitochondrial reads, the background read coverage also increased, resulting in fewer peaks called at the same sequencing depth compared to the original protocol.

Treating the ND samples with anti-mt CRISPR, i.e. combining anti-mt CRISPR treatment with the detergent-free lysis buffer, led to a 3.1-fold decrease in the fraction of mitochondrial reads compared to untreated ND samples (Figure 2.3A, 0.16/0.05). However, unlike DT samples, the fraction of unique, non-mitochondrial reads increased only slightly (Figure 2.3B, median fold-change: 1.1), probably because the fraction of mitochondrial reads was already small. Additionally, the effect of the anti-mt CRISPR treatment was inconsistent, with three samples showing a decrease in the fraction of usable reads and four showing an increase (Figure 2.3B, dashed lines). When calling peaks in anti-mt CRISPR-treated ND samples, this inconsistency was also observed in some of the comparisons performed with different peak calling parameters and read depths, with some of the samples showing an increase in the number of peaks over their untreated counterparts, while other samples had the opposite effect (Supplemental Figure. S2.1).

Figure 2.3 Effect of detergent removal from the ATAC-seq protocol.

(A) The fraction of mitochondrial reads in samples prepared without detergent was considerably smaller than those prepared with the original protocol. Treatment with anti-mt DNA CRISPR led to further decrease of mitochondrial reads (3.1-fold). **(B)** The fraction of unique, non-mitochondrial reads was considerably higher when detergent was not used. Surprisingly, the anti-mt DNA CRISPR treatment had only marginal effect on the fraction of usable reads (1.1-fold increase). **(C)** At the same sequencing depth, only 1.1-fold more peaks were called in the ND treated samples. DT samples 18 and 19 were combined as in Figure 2.2C. **(D)** ND samples displayed higher background (the number of non-mitochondrial reads outside peaks identified in any DT or ND sample). Numbers in blue above the bars are the ratio between number of reads in peaks and the number of background reads (signal/noise) **(E)** An example illustrating the higher background in ND samples, highlighted by the dashed boxes (chr6:420,146-448,555). Fold-differences calculated on medians.

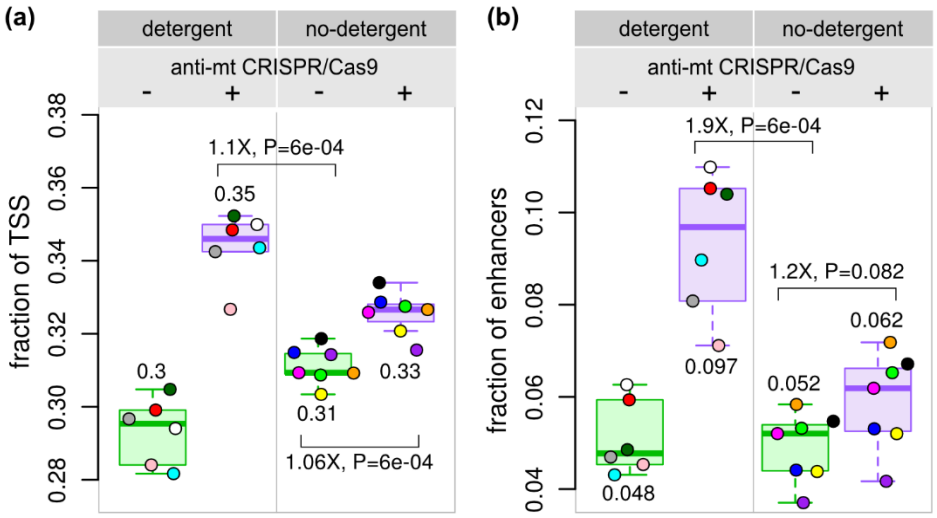


When comparing the effect of the anti-mt CRISPR treatment between ND and DT samples, the former underperformed DT samples in terms of peaks called by 1.3-fold fewer peaks (median of 30,182 vs. 40,682, respectively). Therefore, combining the anti-mt CRISPR treatment with removal of detergent from the lysis buffer did not provide substantial gains over the original protocol with detergent that was treated with anti-mt CRISPR.

In addition to the number of peaks called in the different treatments, we evaluated the quality of peaks using two other parameters: (i) the fraction of Gencode[92] transcription start sites (TSSs) (Figure 2.4A) and (ii) the fraction of Epigenome Roadmap[13] annotated enhancers overlapping peaks (Figure 2.4B and Supplemental Figure S2.2). The highest fraction of TSSs and enhancers was identified in samples treated with anti-mt CRISPR, regardless of whether they were generated with or without detergent. Whereas both anti-mt CRISPR-treated DT and ND samples identified similar numbers of TSSs, enhancers were identified at a higher rate using detergent in conjunction with the anti-mt CRISPR treatment. This difference could be explained by the notion that chromatin tends to be more open in promoters to allow transcription, while enhancers, due to their dynamic nature, would be less accessible. In this scenario, finding enhancers requires lower background and higher quality data, which we have shown is best represented by the detergent anti-mt CRISPR samples.

Figure 2.4 Comparison of the fraction of functional regions overlapping ATAC-seq peaks.

(A) The fraction of transcription start sites (TSSs) overlapping an ATAC-seq peak (+/- 1 kb) was slightly higher in the DT samples than in the ND samples (1.05-fold). **(B)** Treated DT samples identified a greater number of Epigenome Roadmap GM12878 lymphoblastoid cell active enhancers (1.9-fold) than anti-mt CRISPR untreated ND samples. Fold-differences calculated on medians.

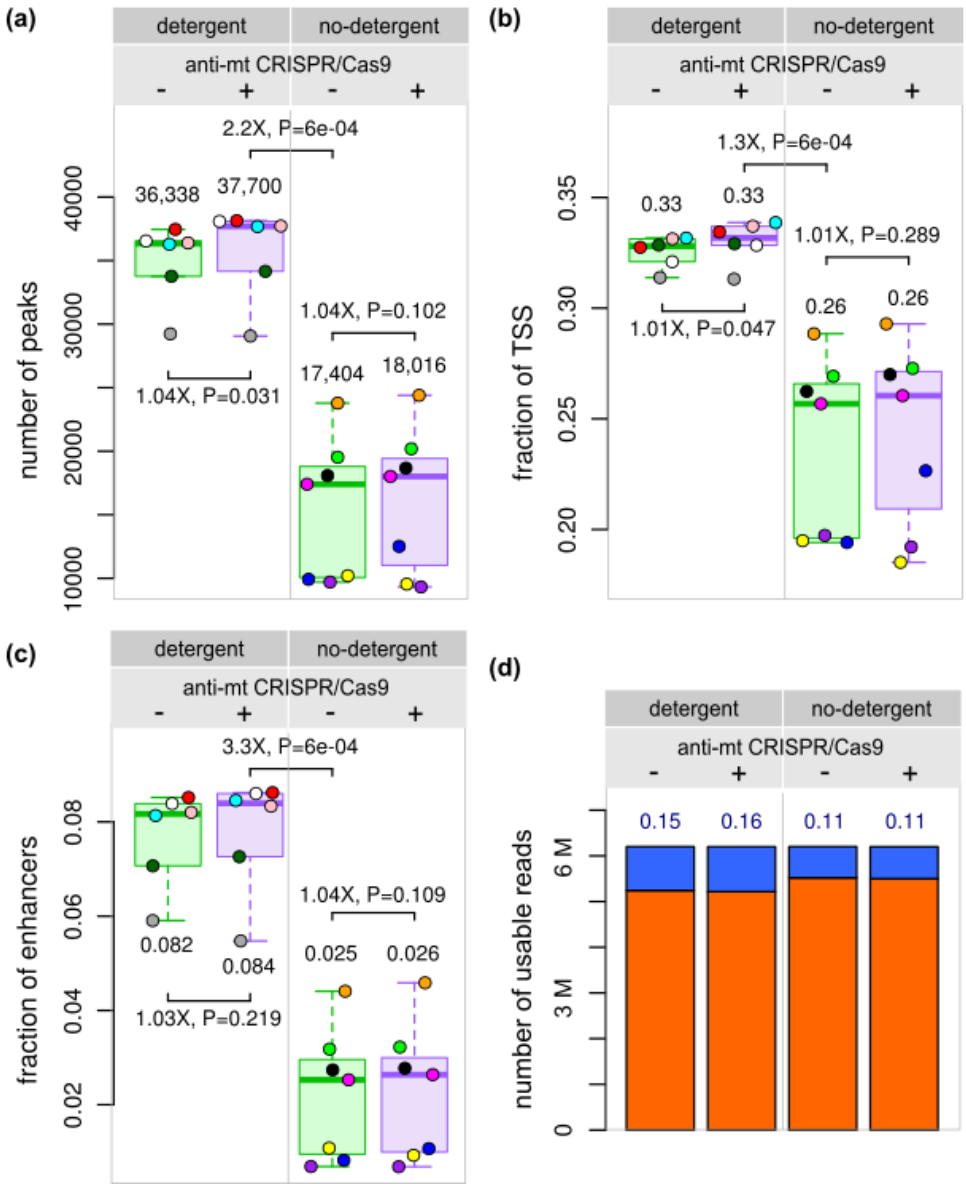


To further investigate differences caused by the anti-mt CRISPR treatment and by removing detergent, we normalized samples by the number of usable reads (Figure 2.5), instead of total sequenced reads (Figures 2.2, 2.3 and 2.4). Figure 2.5 shows that at 6.2 M usable reads, ND samples clearly underperformed DT samples. It also shows that the anti-mt CRISPR treatment removed mitochondrial reads without altering the samples in other ways, since the number of peaks identified, fraction of reads in peaks, fraction of TSS and enhancers identified is the same. Notice that 34 M reads from DT samples (median) were necessary to obtain 6.2 M usable reads, while only 17 M reads from DT samples treated with anti-mt CRISPR were necessary to obtain the same number. We conclude that reducing mitochondrial reads by cleavage of DNA sequencing

fragments using an anti-mt CRISPR strategy yielded the best results in terms of numbers of peaks identified and their quality at the same sequencing depth.

Figure 2.5 Comparison of ATAC-seq samples normalized by total number of usable reads instead of total number of sequenced reads.

(A) The number of peaks is higher in DT than in ND samples. **(B)** The fraction of TSS and **(C)** enhancers identified by ATAC-seq peaks is higher in DT than in ND samples. **(D)** DT samples have more reads in peaks than ND samples. The differences between samples treated with anti-mt CRISPR and left untreated are not statistically significant, showing that the anti-mt CRISPR treatment does not damage the samples.



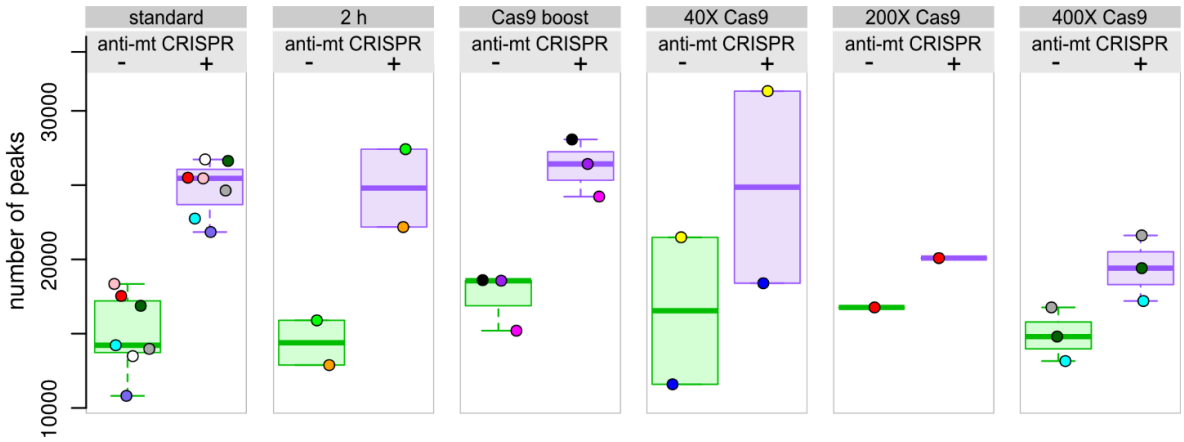
2.3.3 Variations on the anti-mt CRISPR treatment

Given the success of using CRISPR/Cas9 to reduce the amount of mitochondrial reads, we tested modifications of the treatment to enhance the degree of depletion of mitochondrial reads (Figure 2.5). We tested (i) a longer Cas9 incubation of 2 hours instead of 1 hour, (ii) addition of Cas9 for an additional 1 hour after the initial 1 hour treatment (Cas9 boost) and (iii) adding 40X, 200X and 400X Cas9 instead of 100X. None of the treatments led to enhanced depletion of mitochondrial reads and, intriguingly, the 200X and 400X Cas9 treatments performed poorer than the 100X treatment (Figure 2.6).

We did not test a larger number of gRNAs targeting the mitochondrial genome, but it is likely that using 200 gRNAs instead of 100, for example, could further reduce the fraction of mitochondrial reads. However, as guide RNAs are priced per unit, the cost of the treatment increases linearly with the number of targets.

Figure 2.6 Modifications of the anti-mt CRISPR treatment.

Compared to the treatment shown in Figure 2.1 (100X gRNA, 100X Cas9, 1h incubation), labeled "standard", modifications in the treatment did not show improvement. The number of peaks is comparable or even lower in the modified treatments, compared to the standard treatment. Due to the low number of reads in 6 samples, the results presented were obtained with 9.8 M reads randomly sampled. See also Supplemental Figure S2.3.



2.4 Discussion

Our CRISPR/Cas9 treatment targeting 100 loci of the human mitochondrial chromosome successfully reduced the number of mitochondrial reads in LCLs by 1.7-fold, similarly to Wu *et al.*[87], and increased the number of usable reads by 1.6-fold. Consequently, at the same read depth, samples generated with the original ATAC-seq protocol (DT) and treated with CRISPR/Cas9 and anti-mt gRNAs resulted in 1.6-fold more peaks than the untreated controls. More TSSs and enhancers were identified by peaks called in the treated samples, showing that the treatment increases the signal and does not induce unwanted changes in the data.

Removing detergent from the cell lysis step (ND) resulted in even lower number of mitochondrial reads (3.1-fold), but the peaks called were fewer and of lower quality. While the anti-mt CRISPR treatment improved ND samples, resulting in increased number of peaks called, it did not improve over DT samples. We observed more variability in treated ND samples than DT samples, as well as higher background, lower number of peaks and lower overlap with LCL enhancers.

In conclusion, our data show that treating samples prepared using detergent with gRNAs/Cas9 targeting mtDNA was the best way to reduce mtDNA contamination in LCLs, increase the number of peaks, and improve identification of features such as TSSs and enhancers. Given the cost of gRNA oligos, sacrificing sequencing reads may be more economical than depleting mitochondrial reads if only a few samples are generated. In Supplemental File S2.2 we provide a cost calculator based on the numbers obtained in this study and the cost of one lane of sequencing at the University of Chicago Functional Genomic Core Facility. As we have not tested the anti-mt CRISPR treatment and detergent removal in other cell types and cell lines, it is possible that different results may be obtained in other systems, which will affect the cost.

Caution should be taken when multiplexing anti-mt CRISPR-treated samples with samples that have not been treated. Treated samples will yield fewer sequencing reads unless a higher library concentration is used relative to other untreated samples. This is because the cleaved mitochondrial fragments will remain in the library but will not be sequenced since they cannot be amplified by bridge amplification. Sequencing a full lane of samples treated the same way does not require any adjustments.

During the execution of this project, an improved ATAC-seq method was published, termed Fast-ATAC[93], which uses a milder detergent in the cell lysis buffer. This treatment was reported to decrease the fraction of mitochondrial reads from 50% to 11%, while increasing the enrichment of reads in peaks over background and yielding more fragments per cell. The authors noted that cells that are more resistant to lysis may require a stronger detergent, i.e., the original ATAC-seq protocol, in which case, using the CRISPR treatment we analyzed here will remain useful. Since the cost of gRNAs is fixed and can be distributed among multiple laboratories, reducing mtDNA contamination using an anti-mt CRISPR treatment could still lead to significant savings if large numbers of samples are generated.

2.5 Methods

2.5.1 Human lymphoblastoid cell line growth and harvesting

Human lymphoblastoid cell line NA19193 was obtained from Coriell Cell Repository. Cells were grown in RPMI 1640 medium lacking L-Glutamine (Corning), supplemented with 15% fetal bovine serum, 1% GlutaMAX (ThermoFisher) and 1% penicillin-streptomycin solution (ThermoFisher) at a density of $0.5x10^6$ to $1.0x10^6$ cells/mL. Cells were passaged every 2-3 days to maintain this density. Cells were harvested for ATAC-seq by centrifugation at 500 x g for 5

minutes at 4°C and resuspended in PBS. Cells were counted using a hemocytometer and 50,000 cells were immediately placed into a 1.5 mL Eppendorf tube for ATAC-seq.

2.5.2 Preparation of ATAC-seq libraries

ATAC-seq libraries were generated according to the protocol of Buenrostro *et al.*[10] with minor changes. Instead of NEB Next High-Fidelity 2X PCR Master Mix, we used Q5 Hot Start High-Fidelity 2X Master Mix (New England Biolabs). Following PCR-amplification, instead of using a column to clean the reaction, we used a 0.8X Ampure bead purification and eluted the library with 20 μL nuclease-free water. For the ND samples, Igepal-CA630 was removed from the lysis buffer and replaced with water. One microliter of the library was used to run a high sensitivity Bioanalyzer to determine fragment size distribution and concentration.

2.5.3 Anti-mitochondrial CRISPR/Cas9 treatment

To deplete ATAC-seq libraries of DNA fragments derived from the human (hg38) mitochondrial genome, 100 high-quality guide RNAs that specifically targeted the mitochondrial genome roughly every 250 base pairs were chosen using the gRNA design tool at http://crispr.mit.edu (full list of guide sequences is in Supplemental File S2.1). Full-length guide RNAs were designed according to Gu *et al.*[86] and generated from single-stranded oligo templates (Integrated DNA Technologies) according to Lin *et al.*[94]. Briefly, each oligo consisted of the sequence 5'-TAATACGACTCACTATAG(N_{20})GTTTAAGAGCTATGCTGGAAACAGCATAGCAAGTTTAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTTTT-3' where N_{20} corresponds to the 20 nucleotide guide RNA seed sequence. The PAM sequence

would occur at the 3' end of the N_{20}sequence. Oligos were purchased as a 200 picomole plate from Integrated DNA Technologies and received as a lyophilized pool. They were resuspended in 1 mL TE 1.0 buffer (10 mM Tris-HCl, 0.1 mM EDTA). A Nanodrop was used to determine the concentration of the oligos and 8 ng were used as template for PCR to make them double-stranded. The PCR reaction consisted of 4 µL 5X HF Buffer (New England Biolabs), 0.4 µL 10 mM dNTPs, 1 µL of each 10 µM primer (For: 5'-TAATACGACTCACTATAG, Rev: 5'-AAAAAAAGCACCGACTCGGTGC), 0.2 µL Phusion High-Fidelity DNA Polymerase (New England Biolabs) and nuclease-free water to a final volume of 20 µL. Thermocycler conditions were 98°C for 30 s, followed by 30 cycles of 98°C for 10 s, 56°C for 10 s, 72°C for 10 s, and then a final extension of 72°C for 5 minutes. The reaction was cleaned using a Qiagen MinElute Purification kit and eluted in 10 µL of nuclease-free water. Enough PCR reactions were performed to obtain 1 µg of double-stranded template (should be in a volume of less than 8 µL). Transcription was carried out on 1 µg of template using the MEGAshortscript T7 Transcription kit (Thermo Fisher) following manufacturer's instructions and then cleaned with the MEGAclear Transcription Clean-Up kit (Thermo Fisher). gRNAs were eluted from the column with 50 µL of RNase-free water and the concentration was determined using a Nanodrop, aliquoted and stored at -80°C.

We estimated that half of the DNA in each library was of mitochondrial origin, thus a 20 nM ATAC-seq library contained a mtDNA target concentration of 10 nM. Based on this value, 40, 100, 200, or 400 molar excess of Cas9 enzyme was used (New England Biolabs #M0386M) along with 100 molar excess gRNAs in a 30 µL reaction. The reaction was set up according to the protocol for Cas9 from *S. pyogenes* (New England Biolabs #M0386M). Briefly, the appropriate amounts of Cas9 enzyme and gRNAs were mixed with 3 µL of 10X Cas9 Buffer and water to a final volume of 22 µL. This was incubated at 25°C for 10 minutes and then 8 µL of the ATAC-seq

library was added and the reaction was incubated at 37°C for one hour. For the two-hour treatment, the incubation was extended an additional hour; for the "Cas9 boost" treatment, the same amount of Cas9 enzyme was added after 1 hour of incubation and left for an additional hour. Reactions were subsequently treated with 1 μL of 20 mg/mL proteinase K for 15 minutes and purified using a Qiagen MinElute kit followed by elution in 10 μL nuclease-free water. Treated libraries were run on a high sensitivity Bioanalyzer chip to assess fragment size distribution and concentration (Supplemental Figure S2.4). Because the multiplexing barcodes are added before treatment, for each batch of experiments, samples were sequenced on two lanes of an Illumina Hi-Seq 4000 instrument, separating anti-mt-CRISPR untreated and treated samples.

2.5.4 Peak calling

Illumina reads were trimmed using cutadapt[95] and aligned to hg38 with Bowtie 2 version 2.2.3[96] with default parameters. Reads with mapping quality lower than 10 were discarded. Mitochondrial reads and reads aligned to the same coordinates were removed. HOMER version 4.8.3 was run with 3 sets of parameters: (i) "default": -style dnase -gsize 2.5e9, (ii) "ENCODE": -localSize 50000 --size 150 --minDist 50 -fragLength 0 (https://www.encodeproject.org/pipelines/ENCPL035XIO/), (iii) "custom": -gsize 2.5e9 -F 2 -L 2 -fdr 0.005 -region. MACS2 version 2.1.0 was run with 2 sets of parameters: (i) "default": --nomodel --shift -100 --extsize 200 -q 0.01, (ii) "custom": --nomodel --llocal 20000 --shift -100 --extsize 200.

2.5.5 Fraction of TSS and enhancers intersecting peaks

Transcription start sites were obtained from the Gencode[92] GRCh38 basic set (ftp://ftp.sanger.ac.uk/pub/gencode/Gencode_human/release_24/gencode.v24.basic.annotation.gtf.gz), totaling 106,926 2 kb intervals centered on the TSS, and intersected with ATAC-seq peaks using bedtools intersect with the -u option[97]. Epigenome Roadmap[13] 15-state ChromHMM coordinates were obtained from http://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/coreMarks/jointModel/final/all.mnemonics.bedFiles.tgz. Coordinates were converted to hg38 using the UCSC Genome Browser liftOver tool and active enhancer (Enh7) states were intersected with peaks using bedtools with the -u option.

2.5.6 Mean fraction of common peaks and mean Pearson's R^2 of read counts

We ranked peaks called by MACS2 by –log10(qvalue) and used bedtools intersect to count the number of peaks common between the top 20,000 peaks of each sample. The fraction presented in Figure 2.2F is the mean fraction of peaks common between samples of a given group (e.g. treated vs. treated). To calculate the degree of similarity of read counts in ATAC-seq peaks, we merged all peaks from all samples and counted the number of reads in each peak in each sample. We then calculated the $R^2$of the read counts per peak in pairs of samples in each group (e.g. treated vs. treated) and obtained the mean per group presented in Figure 2.2F.

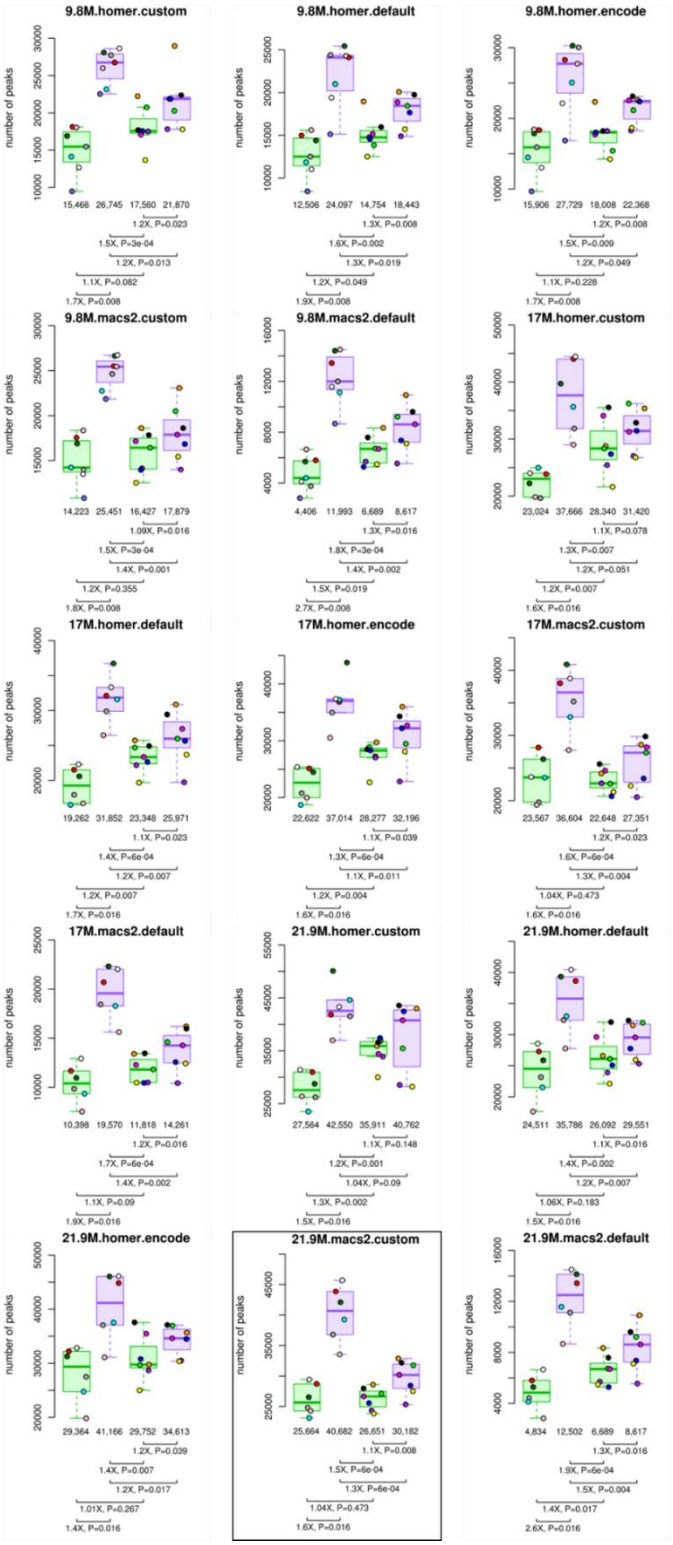
2.5.7 Statistical tests

Due to the small number of replicates, we chose the more conservative Wilcoxon rank sum test to compare treatments in the boxplots shown (R statistical package version 3.3.1). Student t-

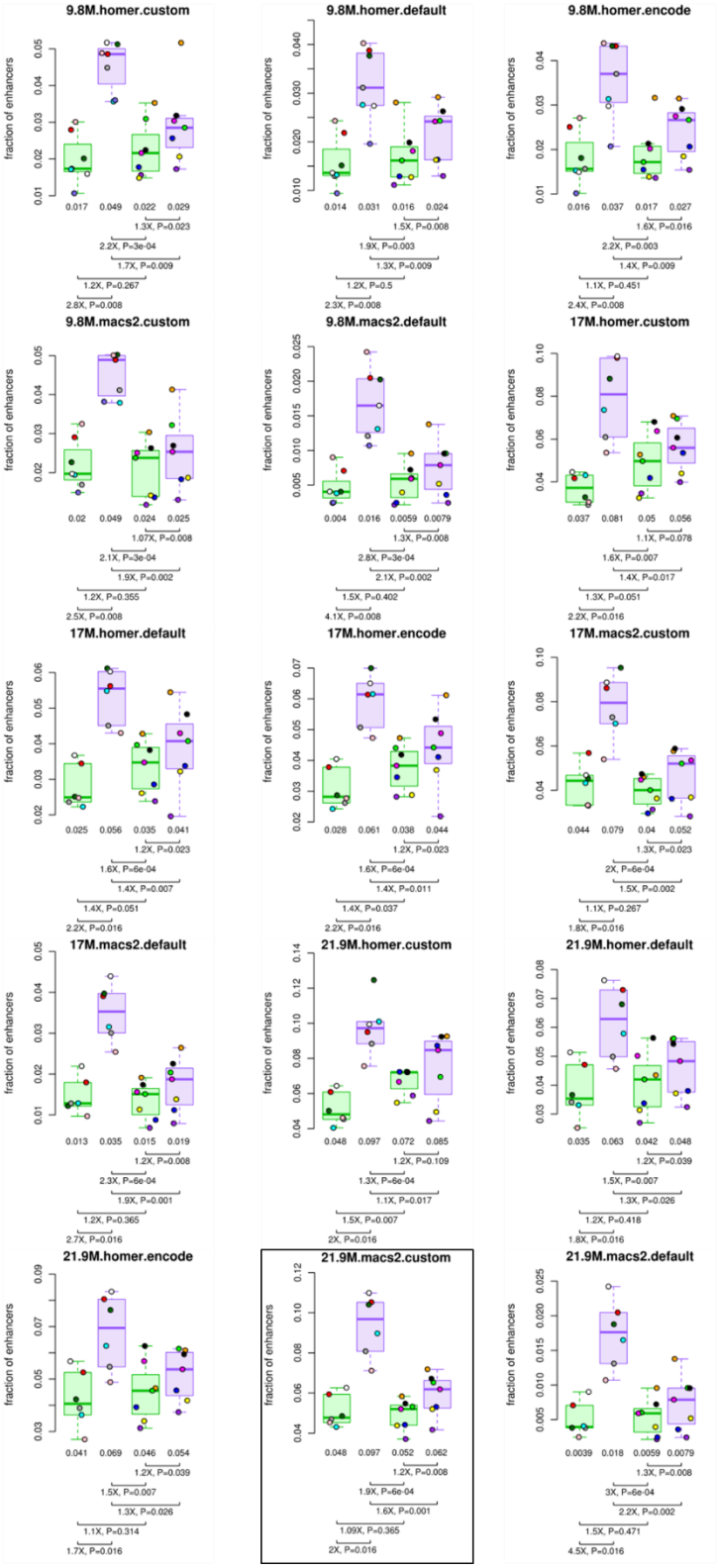
tests were in agreement with the results presented, yielding smaller P-values. Paired tests were used to compare treated/untreated pairs and unpaired tests were used to compare samples prepared with and without detergent. One-sided P-values are presented, since we were interested in specific directions of change. Two-tailed P-values do not change our conclusions. Fold-differences of DT samples were calculated pairwise and the median was reported. Fold-differences of DT versus ND samples were calculated on the median of each group.

2.6 Appendix A: Supplemental Figures

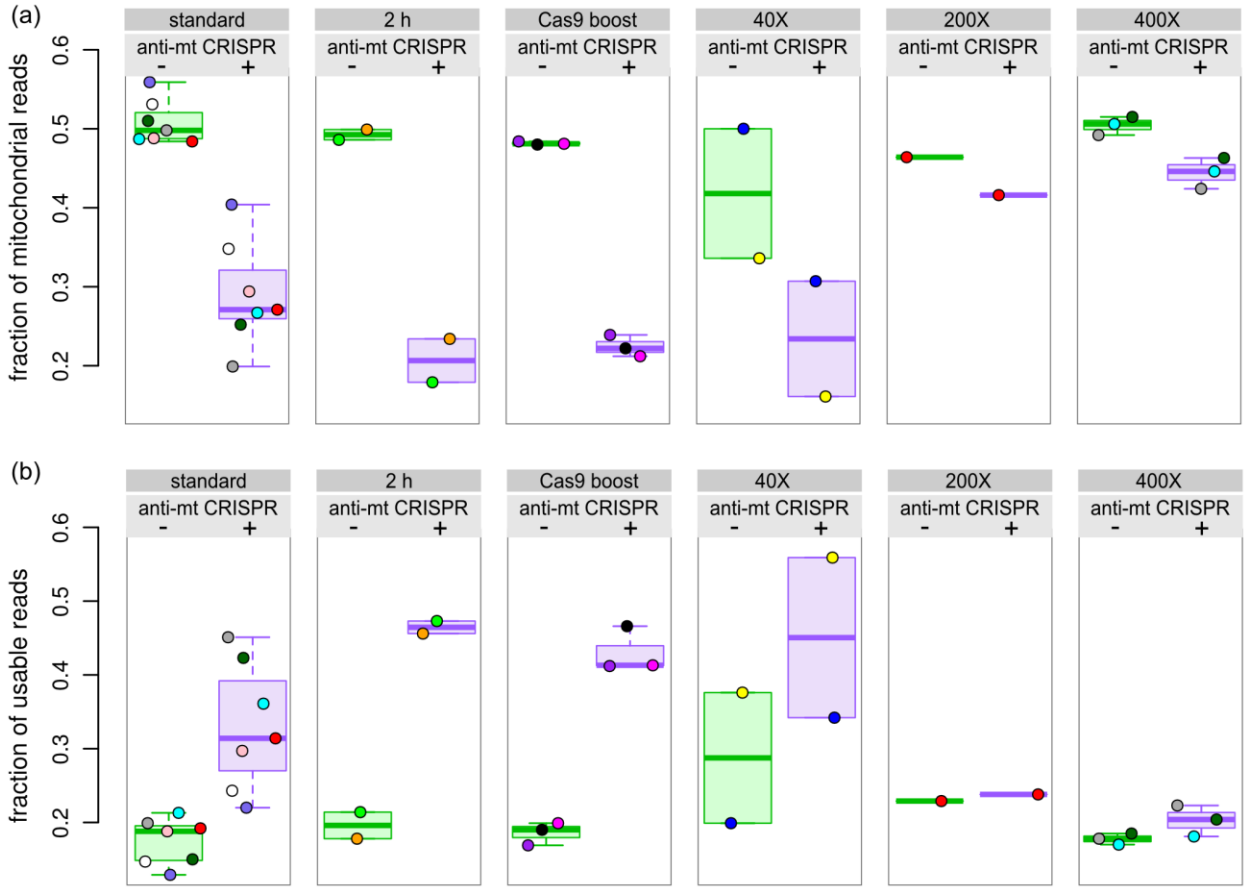
Supplemental Figure S2.1 Number of peaks identified with HOMER and MACS2 using different parameters and 9.8 M, 17 M and 21.9 M reads. Compare to Figure 2.3C of the main text. The plot shown in the main text is shown in a frame.



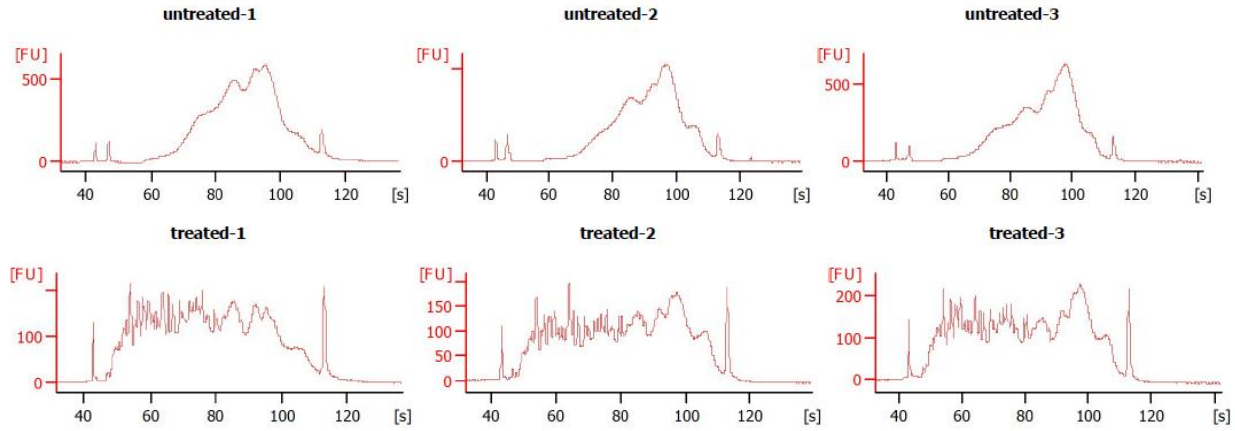
Supplemental Figure S2.2 Fraction of peaks overlapping Epigenome Roadmap lymphoblastoid cell enhancers. Compare to Figure 2.4B of the main text. The plot shown in the main text is shown in a frame.



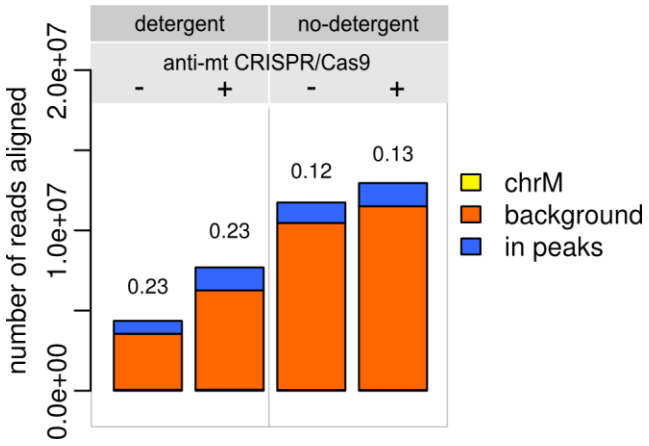
Supplemental Figure S2.3 Effect of modifications of the anti-mt CRISPR/Cas9 treatment on the fraction of mitochondrial reads and usable reads. The number of peaks is shown in Figure 2.5.



Supplemental Figure S2.4 High sensitivity Bionalyzer traces showing 3 replicates of ATAC-seq libraries before and after CRISPR/Cas9 treatment. The X-axis shows the number of seconds that takes the fragments to move through the channel – the longer the fragment, the longer the time. Note the increase in small fragments in the treated samples which theoretically corresponds to cleaved mtDNA sequences.



Supplemental Figure S2.5 Background is higher in ND samples. This figure is similar to Figure 2.3D of the main text but using unique reads. The fraction of chrM reads is too small to be displayed. ND samples have higher background than DT samples.



2.7 Appendix B: Supplemental Tables

Supplemental Table S2.1 Summary of fold-differences between the number of peaks called in anti-mt CRISPR treated (TR) and untreated samples (UN), and samples prepared with (DT) and without detergent (ND) in the cell lysis buffer. Data from Supplemental Figure S2.1.

Analysis	UN-DT x TR-DT	UN-DT x UN-ND	TR-DT X TR-ND	TR-DT X UN-ND	UN-ND x TR-ND
9.8M-homer-custom	1.7	1.1	1.2	1.5	1.2
9.8M-homer-default	1.9	1.2	1.3	1.6	1.3
9.8M-homer-encode	1.7	1.1	1.2	1.5	1.2
9.8M-macs2-custom	1.8	1.2	1.4	1.5	1.09
9.8M-macs2-default	2.7	1.5	1.4	1.8	1.3
17M-homer-custom	1.6	1.2	1.2	1.3	1.1
17M-homer-default	1.7	1.2	1.2	1.4	1.1
17M-homer-encode	1.6	1.2	1.1	1.3	1.1
17M-macs2-custom	1.6	0.96	1.3	1.6	1.2
17M-macs2-default	1.9	1.1	1.4	1.7	1.2
21.9M-homer-custom	1.5	1.3	1.04	1.2	1.1
21.9M-homer-default	1.5	1.06	1.2	1.4	1.1
21.9M-homer-encode	1.4	1.01	1.2	1.4	1.2
21.9M-macs2-custom	1.6	1.04	1.3	1.5	1.1
21.9M-macs2-default	2.6	1.4	1.5	1.9	1.3
mean	1.8	1.2	1.3	1.5	1.2
stdev	0.4	0.1	0.1	0.2	0.1

Supplemental Table S2.2 Fold-differences between the fraction of enhancers identified in anti-mt CRISPR treated (TR) and untreated samples (UN), and samples prepared with (DT) and without detergent (ND) in the cell lysis buffer. Data from Supplemental Figure S2.2.

Analysis	TR-DT X UN-ND	UN-ND x TR-ND
9.8M-homer-custom	2.2	1.3
9.8M-homer-default	1.9	1.5
9.8M-homer-encode	2.2	1.6
9.8M-macs2-custom	2.1	1.07
9.8M-macs2-default	2.8	1.3
17M-homer-custom	1.6	1.1
17M-homer-default	1.6	1.2
17M-homer-encode	1.6	1.2
17M-macs2-custom	2	1.3
17M-macs2-default	2.3	1.2
21.9M-homer-custom	1.3	1.2
21.9M-homer-default	1.5	1.2
21.9M-homer-encode	1.5	1.2
21.9M-macs2-custom	1.9	1.2
21.9M-macs2-default	3	1.3
mean	2	1.3
stdev	0.5	0.1

CHAPTER 3: INVESTIGATING THE ROLE OF ULTRACONSERVED ELEMENTS IN *Irx3* AND *Irx5* REGULATION IN MICE

3.1 Abstract

Ultraconserved elements (UCE) represent the most deeply conserved DNA sequences in mammalian genomes and are thought to function as critical developmental enhancers. However, a complete understanding of the biological functions encoded in these sequences is still lacking. Here, we used genome editing to remove two of the most highly conserved UCEs in the mouse genome, termed UCE3 and UCE5, located near the deeply conserved transcription factor genes *Irx3* and *Irx5*. We show that deletion of UCE5, but not UCE3, results in a significant reduction in body weight, supporting a role for this element as a critical developmental enhancer in *Irx3*/*Irx5* gene function. However, we did not observe any gene expression defects in *Irx3* or *Irx5* in the hypothalamus of UCE5 knock out mice, suggesting regulatory redundancy may exist between the two *Irx* UCEs. When considered together with all available UCE deletion data, our results suggest that the full repertoire of functions encoded in their evolutionarily-constrained sequences likely extends beyond stereotypical enhancer function.

3.2 Introduction

Nucleotide sequence conservation is often used to identify genomic regions that play critical roles in genome function, for example protein-coding sequences and cis-regulatory elements such as enhancers[98]. Ultraconserved elements (UCEs) represent a unique class of 481 DNA sequences which exhibit remarkable sequence conservation of 100% identity over at least 200 base pairs between the human, mouse and rat genomes[22]. Slight relaxation of this constraint identifies hundreds of additional sequences that are nearly identical between the human and fish genomes, representing nearly 500 million years of evolution[23,99]. Because these elements are more

strongly conserved than most protein coding sequences, understanding the vital biological functions maintaining this extreme sequence constraint are of great interest.

A substantial fraction of UCEs are non-coding and are located near genes important in development and transcriptional regulation, suggesting that they may act as critical developmental regulatory elements[22,23,99]. Consistent with this hypothesis, a majority of UCEs tested in *in vivo* reporter assays exhibit tissue-specific enhancer activity during embryonic mouse development[24,100], and many UCEs are transcribed as non-coding RNAs in a variety of human tissues[101]. Furthermore, high resolution mass spectrometry analysis of 193 UCEs revealed that they bind hundreds of transcription factor proteins, with each nucleotide likely contributing to the binding of at least one factor, offering a potential explanation for the extreme nucleotide sequence conservation and tissue-specific enhancer activity[102].

Despite multiple levels of evidence that UCEs function as highly conserved enhancer elements, the three studies to date that have deleted one or more UCEs from the mouse genome have reported surprisingly few or even the absence of molecular or physiological phenotypes resulting from these deletions[26–28]. In the first study, Ahituv *et al*. deleted four different UCEs from the mouse genome—each capable of driving enhancer activity and located near an essential transcription factor gene—yet these deletions did not significantly alter the expression of any nearby gene and the knock-out mice failed to show a phenotype[26]. Subsequently, Nolte *et al*. focused on a UCE that drove reporter gene expression in a specific pattern in the developing limb bud, yet limb development was unaffected after deleting this UCE, although the expression of two nearby genes was decreased[27]. Interestingly, the authors reported that despite the lack of a limb phenotype, mice lacking the UCE weighed approximately 10% less than their wild-type littermates, an effect that persisted into adulthood[27]. Finally, Dickel et al. deleted several UCEs

individually or in combination near the *Arx* gene which plays a role in central nervous system development[28]. The effect of individual deletion of an *Arx* UCE ranged from no phenotype to subtle alterations in the structure of the developing forebrain, marking the first report of a phenotype that mimics what is known about the presumed target gene of the UCE. Importantly, the authors noted that removal of one of the UCEs, termed hs119—which did not result in gene expression or brain development changes—did cause a 10% reduction in body weight, mirroring the result originally reported by Nolte *et al*. Thus, despite displaying strong tissue-specific enhancer activity, most UCEs analyzed in deletion studies appear dispensable for normal development and proper control of gene expression, leaving open the question of the vital functions these DNA elements play in vertebrate biology.

The majority of UCEs lack sequence homology with each other, consistent with their hypothesized role as gene-specific regulatory elements. However, four UCEs located within the two gene deserts containing the *Iroquois* (*Irx*) transcription factor genes share a highly conserved ~300 bp core sequence[103], setting this set of UCEs apart from the other single-occurrence elements. The ancestral *Irx* cluster, which contains *Irx1*, *Irx2*, and *Irx4* on mouse chromosome 13, duplicated nearly 500 million years ago—predating evolution of the major vertebrate lineages—and generated *Irx3*, *Irx5*, and *Irx6* on chromosome 8[104]; remarkably, the positions of the duplicated UCEs relative to the *Irx* genes have been maintained, suggesting that this synteny may be important for their function.

The *Irx* genes encode TALE-class homeodomain transcription factors which are expressed from early development during neural patterning[105] and control the development of numerous tissues including the brain[106], heart[107], lung[108], and kidney[109]. We previously demonstrated that *IRX3* is the functional target of the strongest genetic association with obesity in humans, with *IRX3*

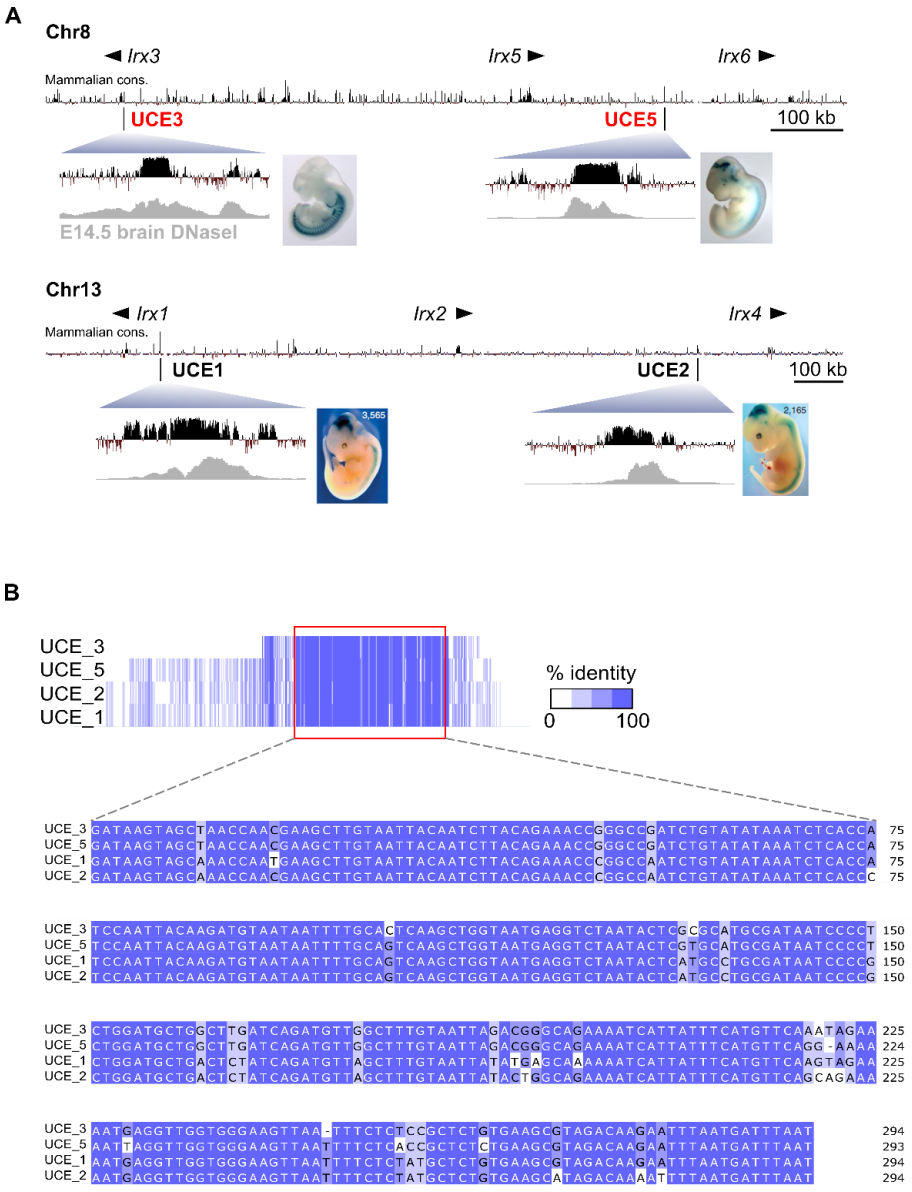
expression in the hypothalamus predictive of obesity risk[74]. Subsequently, both *Irx3* and *Irx5* were identified as important regulators of adipocyte development and thermogenesis in mice, firmly placing *Irx3* and *Irx5* as critical regulators of body weight homeostasis[110]. Because of the uniquely high level of conservation among the UCEs near *Irx3* and *Irx5*, we hypothesized that these elements would be required for proper regulation of *Irx3* and *Irx5*, particularly with respect to maintaining body weight. In this work, we independently deleted these two UCEs from the mouse genome and analyzed the effect of these deletions on body weight and on *Irx3* and *Irx5* expression in the hypothalamus.

3.3 Results

3.3.1 Extreme sequence conservation and functional characterization of the Irx UCEs

The six paralogous *Iroquois* transcription factor genes are located in two ~1 Mbp gene deserts on mouse chromosomes 8 and 13, respectively (Figure 3.1A). Each cluster contains two UCEs—termed UCE1, UCE2, UCE3, and UCE5—which display strong hypersensitivity to DNaseI treatment during embryonic brain development[12], and three of the elements drive enhancer activity in the embryonic midbrain[111] (Figure 3.1A). Unlike the majority of UCEs genome-wide, the four *Irx* UCEs share a 90% identical ~300 bp core sequence (Figure 3.1B). To test our hypothesis that UCE3 and UCE5 act as *cis*-regulatory elements to control proper *Irx3* and *Irx5* expression, we used CRISPR/Cas9 to generate two independent C57BL/6J mouse lines of the UCE3 and UCE5 genomic deletion, respectively. The positions of guide RNAs used to delete each UCE and the respective genotyping strategies are depicted in Supplemental Figures S3.1A,B. Breeding heterozygous animals yielded the expected ratio of offspring of each genotype, as reported previously for other UCEs[26,28].

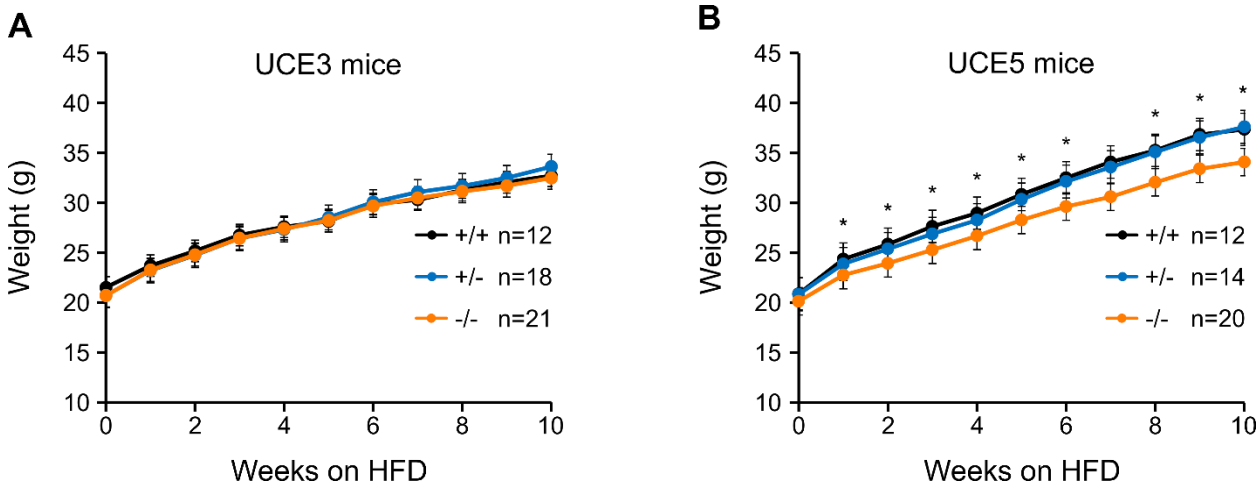
Figure 3.1 Syntenic- and sequence-level conservation of the *Irx* UCEs. (A) Genomic organization and mammalian conservation (PhyloP) of the *Irx* UCEs. The IrxB cluster is shown top, (mm9 coordinates chr8:94,218,409-95,312,733), the IrxA cluster is shown below (mm9 coordinates chr13:72,754,894-73,755,393). Below each UCE is the track displaying DNaseI hypersensitivity data from embryonic day 14.5 (E14.5) brain[12]. To the right of each UCE is a representative E11.5 embryo showing enhancer activity of each UCE (UCE3 and UCE5 from the VISTA Enhancer Browser[111], UCE1 and UCE2 from[112]). UCEs in red were deleted in this study. **(B)** Nucleotide sequence alignment of the four mouse *Irx* UCE elements.



3.3.2 Deletion of UCE5, but not UCE3, results in reduced body weight on a high fat diet

We previously demonstrated that deletion of *Irx3* specifically in the hypothalamus led to reduced body weight in mice fed a high-fat diet[74]. To test the hypothesis that deletion of the UCE3 and UCE5 elements would disrupt metabolic functions of the *Irx3* and/or *Irx5* genes, we placed WT, heterozygous and homozygous UCE deletion animals on a 55% fat diet for 10 weeks and measured body-weight weekly (Figure 3.2). Deletion of UCE3 did not lead to a significant difference in weight (Figure 3.2A). In contrast, homozygous deletion of UCE5 caused a significant ~8% reduction in body-weight compared to both WT and heterozygous littermates (Figure 3.2B), mirroring the metabolic phenotype reported for the hypothalamic-specific *Irx3* knock out animal[74].

Figure 3.2 Effect of UCE3 and UCE5 deletion on body weight. Wild-type, heterozygous, and homozygous littermate UCE3 **(A)** and UCE5 **(B)** animals were placed on a 55% fat diet starting at 6 weeks of age and body weight was measured weekly for 10 weeks. *p < 0.05, between wild-type and homozygous groups, two-tailed t test.

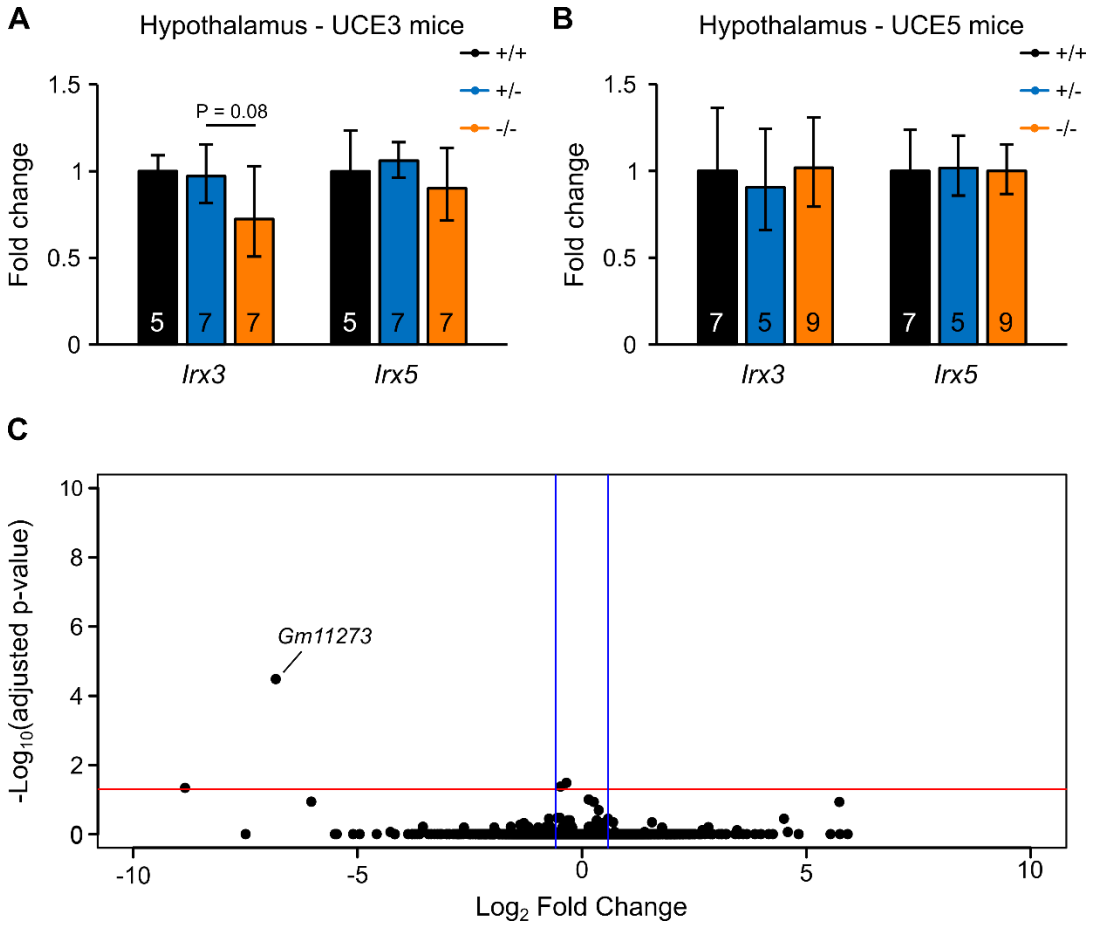


3.3.3 UCE3 and UCE5 are not required for Irx3 or Irx5 expression in the adult hypothalamus.

We expected that deletion of UCE5 would lead to decreased hypothalamic expression of *Irx3* or *Irx5* based on the observed body weight phenotype and consistent with previous results of *Irx3* and *Irx5* function[74,110]. We harvested the hypothalamus from animals after they had completed the ten-week high fat diet regimen, and measured expression of *Irx3* and *Irx5* using quantitative real-time PCR (qPCR) (Figure 3.3). As expected, we did not observe any significant difference in *Irx3* or *Irx5* expression in the hypothalamus of UCE3-/- animals compared to wild-type, consistent with the lack of a body weight phenotype (Figure 3.3A). We note that the data suggest *Irx3* expression may be influenced by the UCE3 deletion, although the result did not reach statistical significance (Figure 3.3A). Surprisingly, we also did not observe any significant difference in *Irx3* or *Irx5* expression for the UCE5-/- animals, despite the significant body weight phenotype (Figure 3.3B).

To rule out the possibility that the UCE5 deletion caused a body weight phenotype through action on other genes, either directly in *cis* or indirectly through modifications in *Irx3* or *Irx5* expression that were not detected via qPCR, we next performed RNA-seq on five wild-type and five UCE5-/- hypothalamus samples. Consistent with the qPCR results, RNA-seq revealed that the gene expression profiles of wild-type and UCE5 deletion samples are remarkably similar, with only one gene identified as significantly differentially expressed (Log2 fold change > 1.5, adjusted p-value < 0.05, Figure 3.3C). This gene is *Gm112873*, a predicted pseudogene derived from *Cox5b* which encodes a subunit of the cytochrome C oxidase complex. In summary, we find no evidence that adult hypothalamic gene expression is perturbed in UCE5-/- mice despite their significant body weight phenotype.

Figure 3.3. Gene expression analysis in UCE3 and UCE5 deletion mice. (**A** and **B**) qPCR analysis of *Irx3* and *Irx5* in the hypothalamus of wild-type, heterozygous and homozygous UCE3 **(A)** and UCE5 **(B)** deletion animals after finishing the high fat diet regimen. Sample sizes are indicated within the representative bars. Data are presented as mean ± s.e.m. **(C)** RNA-seq analysis for differentially expressed genes using DEseq2 in 5 wild-type and 5 homozygous UCE5 deletion hypothalamus samples. Each dot represents a gene. Genes falling above the red line (fold change > 1.5) and to the left or right of the vertical blue lines (adjusted p-value < 0.05) are significant. Using this threshold, only *Gm11273* is statistically differentially expressed.



3.4 Discussion

Since their original discovery and characterization fifteen years ago, ultraconserved elements have been extensively studied in an effort to understand the biological explanation for this extreme level of conservation. Despite a multitude of functional[24,100–102], computational[113,114], and genetic deletion studies[26–28], a clear understanding of their vital biological function is still lacking. Most studies involving functional or computational approaches were able to analyze large numbers of elements, whereas deletion studies are limited in throughput by the inherent time- and cost-constrained nature of generating and phenotyping transgenic animals. We sought to add to the small yet growing amount of data on UCE deletion mouse models in order to increase our understanding of the genetic and physiological impact of removing these elements from the genome.

A role for UCE5 in Irx3/Irx5-regulated body weight homeostasis

We focused on two UCEs, UCE3 and UCE5, that are among the most highly conserved sequences in the human genome as these non-coding elements have retained a 90% identical ~300 bp core sequence over the last 500 million years of vertebrate evolution. Based on this unique level of conservation, the regulatory characteristics of these elements during embryonic brain development, and previous work that elucidated roles for *Irx3* and *Irx5* in body weight homeostasis[74,110], we predicted that deletion of UCE3 or UCE5 individually from the mouse genome would disrupt normal *Irx3* and/or *Irx5* gene expression. We previously established that hypothalamic-specific knock-out of *Irx3* causes mice to gain less weight when fed a high fat diet compared to animals with normal *Irx3* expression[74]. We therefore predicted that UCE3-/- and/or UCE5-/- mice would similarly show a body weight phenotype when fed a high fat diet, and that this effect would be driven by hypothalamic changes in *Irx3*/*Irx5* expression.

Our results show that deletion of UCE5, but not UCE3, does indeed lead to reduced weight gain when fed a high fat diet, mimicking the phenotype observed for hypothalamic *Irx3* knock-out animals. Despite this similarity, we did not observe any change in *Irx3* or *Irx5* expression in the hypothalamus of these animals. Furthermore, RNA-seq revealed that only one gene was significantly differentially expressed between UCE5-/- and wild type hypothalamus samples, indicating that the gene expression profile in the hypothalamus of UCE5-/- animals is not different from wild-type. However, because we only included five samples in each group, we cannot rule out that subtle changes in gene expression may drive the body weight phenotype.

There are several explanations that may reconcile our observation of a body weight phenotype in the absence of a molecular phenotype. First, the time-frame in which UCE5 actively contributed to *Irx3* or *Irx5* expression may have been missed as we only measured *Irx3*/*Irx5* expression in 16 week-old adult mice, after they completed the high fat diet regimen. Data from ENCODE show that UCE5 is highly sensitive to DNaseI activity in E14.5 embryonic brain[12], and UCE5 drives strong reporter gene expression in the developing midbrain of E11.5 embryos[111]. Thus, it is likely that UCE5 acts during brain development and may no longer be active in the adult hypothalamus. If this is the case, the body weight phenotype we observed during adulthood may have been initiated early in development, when the absence of UCE5 would have impacted *Irx3*/*Irx5* expression. As measured in the adult, the expression levels of these genes are the same compared to wild-type animals, indicating that UCE5 is not required for maintenance of *Irx3*/*Irx5* gene expression over time. Similarly, the RNA-seq data show that any changes in gene expression early in development are not retained in the adult hypothalamus, again supporting that UCE5 may act exclusively during brain development, and that gene expression effects are no longer detectable in fully developed hypothalamic tissue.

An alternative explanation is that regulatory redundancy ensures proper *Irx3*/*Irx5* expression in the adult hypothalamus. Because we did not combinatorially delete UCE3 and UCE5 in cis, the presence of the remaining element may have compensated for the deleted element, resulting in the observed wild-type levels of gene expression, as reported previously for numerous limb enhancers[115]. Indeed, Dickel *et al*. found that deletion of two UCEs was often required to observe changes in target gene expression and a corresponding defect in neuron development[28]. The fact that we did not observe a change in *Irx3/Irx5* expression in the adult hypothalamus of UCE5-/- may indicate that UCE3 compensated for this deletion. Alternatively, in light of the hypothesis presented above—that UCE5 acts predominately during embryonic brain development—embryonic *Irx3*/*Irx5* expression in UCE5-/- animals may not have been adequately compensated for by the remaining UCE3 element during this critical developmental window. In other words, *Irx3*/*Irx5* expression may be differentially sensitive to the action of cis-regulatory elements at different stages of development[116].

Our observation that deletion of the UCE3 element did not cause a body weight or molecular phenotype indicates that this element may be more efficiently compensated for by UCE5 at all stages of pre- and post-natal development. Interestingly, UCE3 is the shortest of the four *Irx* UCEs, essentially retaining just the core 300 bp sequence, indicating that this element may be under relatively weaker selection pressure compared to the other elements and consistent with the stronger phenotypic impact we observed with the UCE5 deletion. Additionally, UCE3 is the only *Irx* UCE that did not drive reporter gene expression in E11.5 midbrain[111,112], again supporting our observation that deletion of UCE5, but not UCE3, caused a body weight phenotype. Taken together, our results support a role for UCE5 in regulating body weight through the action of *Irx3*

and/or *Irx5* gene expression in the brain, however we acknowledge the limitation of our work in that we did not directly observe an impact on gene expression in UCE5-/- animals.

Implications of a body weight phenotype on interpretation of function

To date, ten distinct UCEs have been individually deleted from the mouse genome, including those presented in this work (refs[26–28], note that one of the elements deleted in Ahituv *et al.* ("uc467") was also deleted in Dickel *et al.* (termed "hs121")). Only three of the ten deletions resulted in a phenotype related to the known function of the presumed target gene, namely brain development in the case of hs121 and hs122 (target gene *Arx*, Dickel *et al*.) and body weight in the case of UCE5 (target gene *Irx3*/*Irx5*). Interestingly, we noted that two of the UCE deletions that failed to cause an "expected" phenotype nevertheless resulted in decreased body weight: the M280 deletion predicted to impact limb expression[27], and the hs119 deletion predicted to impact neuron development[28]. When considered with our present work, the total number of UCEs that lead to a body weight phenotype when deleted individually is 3/10 (30%).

Body weight and body size are important indicators of general organismal fitness[117]. Changes in body weight or size can indicate a spectrum of underlying causes, and are not necessarily a direct result of specific perturbations to metabolic homeostasis. Indeed, in a survey of available gene knock-out mouse strains, it was found that one-third of gene knock-outs cause a body weight phenotype, with the vast majority leading to decreased body weight[74,118]. We were intrigued by the observation that a similar proportion of UCEs cause a reduced body weight when deleted in the mouse, although the sample size is admittedly very small (3/10). To assess whether this is a more general phenomenon whereby a body weight phenotype resulting from a non-coding enhancer knock-out may be indicative of general organismal fitness, we conducted a literature search of all studies that have deleted an enhancer from the mouse germline and measured post-

natal body weight. Surprisingly, we found that 9/33 (27.3%) of individual enhancer deletions caused a body weight phenotype, often in addition to other more specific phenotypes related to the function of the enhancer's target gene (Supplemental Table S3.3. The majority of enhancer deletion studies that we considered did not report body weight measurements, making it difficult to judge how representative this value is of all enhancer deletions. However, the result stands that a substantial proportion of non-coding enhancer deletions cause a body weight phenotype, including three UCEs tested to date.

Because significant changes in body weight indicate an underlying impact on general health and fitness, we may assume that this metric represents a valid readout for biological function of the deleted element. In the case of the UCE5 deletion, we predict that the mode of action is through modification of *Irx3*/*Irx5* expression during embryonic brain development, potentially influencing development of the hypothalamus. This may lead to the body weight phenotype via the same mechanisms as disrupting hypothalamic *Irx3* expression, namely by modifying metabolic homeostasis[74], or it may reflect a more general response to a genetic perturbation during a critical window of development. Lacking other phenotypes or molecular signatures, it is difficult to pin-point the mechanism by which UCE5 deletion causes a body weight phenotype. However, given that two previous UCE deletions hypothesized to influence limb (M280,[27]) and neuron development (hs119,[28]), respectively, failed to cause these phenotypes and rather caused a body weight phenotype suggests that we may only be observing the end-product of some fundamental yet still largely enigmatic function carried out by ultraconserved sequences.

In conclusion, we have shown that deletion of one of the most highly conserved sequences in the mouse genome, UCE5, impacts organismal fitness by reducing body weight. A second deletion of the UCE3 element, which is nearly identical to the UCE5 element in its core sequence,

did not cause a body weight phenotype, arguing that the *Irx* UCEs are not functionally identical despite their highly conserved core sequence. When considered in light of all UCE deletions analyzed to date—the majority of which do not cause gene expression or developmental defects—our results support the notion that although UCEs show characteristic signs of acting as developmental enhancers, the full repertoire of functions encoded in their evolutionarily-constrained sequences likely extends beyond stereotypical enhancer function.

3.5 Methods

3.5.1 Generation of UCE deletion mice

UCE deletion mice were generated with in vivo CRISPR/Cas9 editing as follows. The genomic positions of UCE3 and UCE5 were obtained from[22] and expanded to include the entire conserved block (UCE3: mm9 coordinates chr8:94328725-94329151; UCE5: mm9 coordinates chr8:95069383-95070118). Two guide RNAs (gRNAs) were designed to target the 5' and 3' end of each UCE, for a total of 4 gRNAs per UCE (Supplemental Table S3.1). gRNAs were cloned into the T7cas9sgRNA2 vector (Addgene) according to[119]: gRNAs were ordered as complementary DNA oligos from Integrated DNA Technologies in the form 5'TAGG$N_{gRNA_forward}$ and 5'AAAC$N_{gRNA_reverse}$, where N_{gRNA} corresponds to the gRNA seed sequence found in Supplemental Table S3.1. Oligos were resuspended in water to a final concentration of 100μM and 2 ul of each oligo (forward and reverse) were mixed with 2 ul 10X NEBuffer 3 (NEB B7003S) and 14 ul water. Oligos were annealed by incubating the mixture at 95°C for 5 minutes, ramping down to 50°C at 0.1C/second, incubating at 50°C for 10 minutes and then holding at 4°C. Annealed gRNAs were cloned into the T7cas9sgRNA2 vector by mixing 1 ul annealed oligos, 0.4ug vector,

1 ul 10X NEBuffer 3, 1 ul 10X T4 DNA Ligase Buffer (NEB B0202S), 0.5 ul T4 DNA Ligase (NEB M0202S), 0.5 ul BsmBI (NEB R0580S), 0.3 ul BglII (NEB R0144S), 0.3 ul SalI (R0138S), and water to a final volume of 10 ul. The ligation reaction was carried out in a thermocycler with the following conditions: 3 cycles of 20 minutes at 37°C/15 minutes at 16°C, followed by 10 minutes at 37°C, 15 minutes at 55°C, and 15 minutes at 80°C, and a final hold at 4°C. One ul of the ligation reaction was used to transform 15 ul of chemically compentent *Escherichia coli* cells (Thermo Fisher C404006), which were selected on Ampicillin plates. Colonies were screened by PCR and Sanger sequencing for the correct gRNA insertion. gRNA and Cas9 mRNA transcription reactions were carried out according to[119]. The Cas9 expression vector was ordered from Addgene (47929). For each UCE target, 1.375 ug of each of the four gRNAs was mixed with 16.5 ug of Cas9 mRNA in a final volume of 110 ul. This mixture was injected into the cytoplasm of mouse embryos using standard procedures and implanted into recipient mice. The resulting founder pups were genotyped with primers listed in Supplemental Table S3.1 and a single founder line was established for each UCE deletion. All animals used in this study were obtained by mating heterozygotes.

3.5.2 High fat diet and body weight measurements

Wild-type, heterozygote and homozygote littermates were fed a 55% high-fat diet (Research Diets) starting at 6 weeks of age. All three genotypes were present in each cage. Body weight was measured weekly for 10 weeks. Animals were excluded from the final data set if their weight fell beyond two standard deviations from the mean at five or more time points.

3.5.3 Quantitative real-time PCR

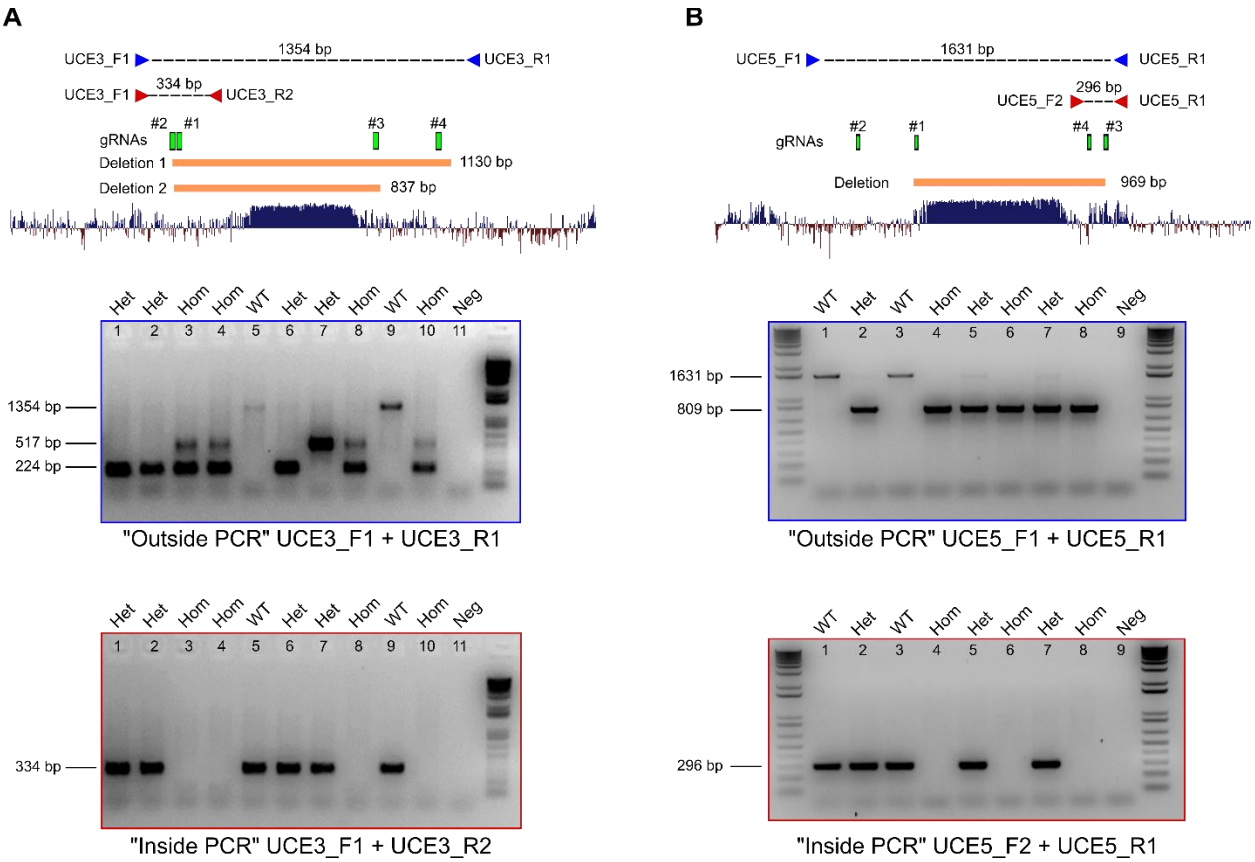
The hypothalamus was harvested from mice after they had completed the high-fat diet regimen. Mice were humanely euthanized and the brain was removed. A coronal section containing the hypothalamus was removed from the brain using an adult mouse brain matrix. The hypothalamus was dissected from this section and immediately flash-frozen in liquid nitrogen until all samples had been collected. To isolate RNA, 300 ul TRI Reagent (Sigma T9424) was added to the frozen tissue and homogenized. Total RNA was isolated with the Zymo Direct-zol RNA purification kit (R2060) following manufacturer's instructions. Complementary DNA was synthesized from 1 ug of RNA using Superscript II Reverse Transcriptase (Thermo Fisher 18064014) with oligo(dT). qPCR was carried out in triplicate reactions with SYBR Green Master Mix (Bio-Rad 172-5270) on a Bio-Rad CFX Connect machine and relative cycle threshold (CT) values were normalized by *Gapdh*. Primer sequences are provided in Supplemental Table S3.1.

3.5.4 RNA-seq

RNA-seq libraries were prepared from a subset of the RNA samples used for qPCR analysis (five wild-type and five UCE5-/-, based on sufficient material available, see Supplemental Table S3.2 for the specific animals). RNA-seq libraries were generated with the Illumina TruSeq V2 kit (Illumina, RS-122-2001) and 500 ng of RNA, following manufacturer's instructions, and sequenced on an Illumina HiSeq 4000. Gene counts were quantified with Salmon 0.9.1[120] and imported with tximport 1.2.0[121] into DESeq2 1.14.1[122] to call differentially expressed genes. A minimum 1.5-fold-difference between WT and UCE5-/- and a minimum adjusted p-value of 0.05 were required to call differentially expressed genes. TPMs (transcripts per million) were also estimated by Salmon.

Appendix C: Supplemental Figures

Supplemental Figure S3.1. Generation of UCE3 and UCE5 knock-out mice. (A) gRNAs used to generate the UCE3 deletion are shown in green and the resulting germline deletions are shown in orange, overlaid on the UCE element. The genotyping strategy is shown with the two sets of primers used to detect the WT (blue and red primer pairs) or deletion band (blue primer pair). A representative PCR result is shown below which identifies WT, heterozygous and homozygous individuals. **(B)** Same as in **(A)** but for the UCE5 deletion.



Appendix D: Supplemental Tables

Supplemental Table S3.1 List of oligos used in the study.

Reagent	Name	Purpose	Sequence
gRNA	UCE3_gRNA1	UCE3 CRISPR	5’GCGCACGAAGCTGTGCGCGC
gRNA	UCE3_gRNA2	UCE3 CRISPR	5’TACAGAAAAGGTGTACGCCG
gRNA	UCE3_gRNA3	UCE3 CRISPR	5’ATTCCTCCCGGTGCTACGCC
gRNA	UCE3_gRNA4	UCE3 CRISPR	5’GGATTTGTACACCGCCGAAC
gRNA	UCE5_gRNA1	UCE5 CRISPR	5’GGCGCCCGTTAAAGCCCTCT
gRNA	UCE5_gRNA2	UCE5 CRISPR	5’GCATTATCTACTGCCTACGG
gRNA	UCE5_gRNA3	UCE5 CRISPR	5’TTATATTTCTGCGATCCGCA
gRNA	UCE5_gRNA4	UCE5 CRISPR	5’GCCGGTTTGTCAAGCTTGGC
genotyping primer	UCE3_F1	genotype UCE3 deletion	5'GACCTGAGCGCAACAGCA
genotyping primer	UCE3_R1	genotype UCE3 deletion	5'CATGCTGAGATTCCGGGAG
genotyping primer	UCE3_R2	genotype UCE3 deletion	5'AGACACAGCTGCAGAAGCT
genotyping primer	UCE5_F1	genotype UCE5 deletion	5'GTGAAGAGCTGGTAAGATCAAGG
genotyping primer	UCE5_F2	genotype UCE5 deletion	5'GCTGTGGCTGGAATGATCTTG
genotyping primer	UCE5_R1	genotype UCE5 deletion	5'GACCTTGTGTGAGGCTCATTG
qPCR primer	mIrx3_F	qPCR for Irx3	5’CAATGTGCTTTCATCAGTGTACG
qPCR primer	mIrx3_R	qPCR for Irx3	5’GGATGCTGGACGCCAGGGCTGT
qPCR primer	mIrx5_F	qPCR for Irx5	5’ACAACTCGCACCTCCAGTACG
qPCR primer	mIrx5_R	qPCR for Irx5	5’CCATAAGGATAGGAGCCCAG

Supplemental Table S3.2 Raw high fat diet data for all animals. Animals were excluded from the analysis if their body weight fell outside of two standard deviations of the mean at 5 or more time-points (animals highlighted in red). Animals from which hypothalamus samples were used in RNA-seq are highlighted in blue (UCE5 only).

Week	0	1	2	3	4	5	6	7	8	9	10
						UCE3 +/+					
415	17.35	20.8	22.9	24.6	26	26.98	29.18	28.22	30.1	31.2	31.92
439	19.5	22.7	24.5	26.38	28.28	26.67	30.5	29.18	31.65	32.04	33.09
476	20.7	22.8	24.2	25.6	27.68	27.76	30.23	31.76	33.19	32.68	34.77
639	22.96	26.25	27.62	30.64	32.24	32.28	33.85	34.5	35.93	36.6	35
694	N/A	25.3	27.54	28.5	27.5	28.9	29.6	30.2	31.9	33.9	35.9
697	N/A	21.6	23.76	25.5	24.3	25.9	28.2	29.3	31.1	32.1	33.6
777	20.9	24	25	28.2	29.6	29.1	31.3	31.6	32.5	32.7	33.4
812	20.6	20.9	21.8	22.4	22.8	24.2	24.4	25.6	25.5	26.2	26.3
813	22.4	23.6	24.4	25.5	26.2	28	28.2	30.5	30.4	32.5	33.1
863	23.9	26	27.6	28.5	30	30.6	32.2	32.2	32.2	32.5	32.5
864	25.5	27.3	28.5	29.3	30.8	31.4	33	33.1	32.3	32.4	31.3
835	21.4	23.2	24.3	26.1	25.7	26	27.7	27.6	28.4	29.8	31.6
						UCE3 +/-					
474	23	24.3	26.5	28.65	29.88	30.68	32.9	33.92	35.95	35.43	36.93
475	23.75	24.3	26.6	28.59	30.18	29.8	31.98	33.2	34.04	34.34	36.41
501	19.6	21.95	23.38	24.12	24.82	27.2	28.2	30.2	31.75	30.4	30.45
502	17.5	20.35	21.62	22.49	23.34	24.44	25.4	26.9	27.63	26.7	26.33
517	21.62	25.61	26.9	28.34	29.51	31.24	33.32	34.66	33.25	32.94	34.03
519	20.88	25.7	27.53	29.12	31.48	32.9	36.9	39.52	37.52	39.78	42.6
548	21.78	22.7	24.8	27.68	28.21	29.73	31.45	32.6	33.28	34.8	36.81
550	20.25	21.78	23.46	26.15	26.38	27.43	29.8	29.9	31.86	33.1	34.6
549	20.45	22.4	23.8	26.11	26	27.48	28.86	29.55	29.78	31.1	33.02
657	21.09	24.56	26.7	27.03	28.92	29.5	30.44	30.9	30.5	31.3	31.6
659	19.56	23.15	24.8	26.25	27.26	27.8	29.45	30.3	30.6	31.5	32.8
690	N/A	22.9	24.38	26.5	25.9	27.1	27.9	29.5	30.4	32	32.4
691	N/A	23.5	25.03	26.5	26.4	27.5	28.3	29.4	29.4	30.8	31.5
773	21.1	24.1	25.4	28	29.6	31.4	34.1	34.5	36.4	36.7	37.9
774	22.8	25.2	27	29.2	30.5	31.8	33.9	34.4	35.9	35.7	36.3
814	22.1	24.1	25.2	26	26.4	27.4	28	29.6	29.6	31.4	31.9
837	17	19.7	20	21.7	22.5	23.4	23.7	23.9	25	27	27.7
838	19	21.6	22.3	23.5	24.5	26.7	26.6	26.4	27.6	30.1	31.9
						UCE3 -/-					

Supplemental Table S3.2, continued

440	19	21.4	23.6	25.7	27.6	28.79	30.9	29.21	31.52	33.42	35.44
409	17.5	21.75	23.8	25.69	27.3	28.3	28.94	28.23	30.7	31.63	32.01
473	23.3	25.8	28.2	30.2	32.2	31.9	34.15	34.85	36.68	36.37	38.09
503	19.16	21.9	22.79	23.65	24.03	25.5	26.2	28.62	30	28.7	27.88
518	20.68	22.9	24.38	25.39	26.8	27.74	30.02	31.59	31.04	30.32	31.61
520	20.7	24.6	26.95	28.21	30.63	31.58	34.78	36.63	34.17	33.22	35.61
547	21.79	23.51	25.54	28.71	28.15	28.93	30.53	31.23	32.3	33.7	35.44
551	21.45	23.57	25.3	28.14	27.55	28.9	30.87	32.42	33.54	34.1	36.86
634	17.26	19.57	20.43	22.5	23.83	24.51	25.3	25.2	24.68	25.3	25.4
637	20.9	23.4	24.2	26.53	27.37	27.68	28.78	28.5	29.7	30.8	29.1
638	21.6	25.13	26.31	28.66	31.6	32.29	32.96	32.9	35.4	36.7	36.2
660	20.55	23.77	25.98	26.69	28.61	28.8	30.04	31.3	31.1	32.2	32.6
658	22	25.03	27.36	28.25	30.07	30.3	32.9	33.9	32.7	33.1	34.5
692	N/A	23.9	25.37	26.9	27.1	27.8	28.2	29.9	29.4	30.5	30.9
693	N/A	22.9	23.92	25.5	25.8	26.9	28	29.6	29.7	30.8	32
695	N/A	22.6	23.56	24.7	24.5	25	26.4	28	29.7	30	31.7
696	N/A	23	24.86	26	25	26	27	27.7	28	28.8	29.7
776	23.3	26.5	28	31.3	32.4	32.8	35	36.4	37.7	37.8	38.6
815	21.8	22.6	23.4	23.5	24.4	25.9	26	27.6	27.5	28.8	30.2
861	22.5	24.3	26	26.6	28	28.7	29.9	30.2	30.9	31.5	30.8
836	18.4	20.3	20.7	22.4	22.8	24.3	26.4	26.1	27.5	28	27.6

UCE5 +/+

450	18.9	21.9	24.1	25.4	26.95	28.76	29.4	29.1	29.5	31.2	31.66
453	20.2	23.75	26.6	28.7	30.5	33	34.36	35.44	35.5	38.8	39.74
840	21.5	24.15	25.6	28.45	28.3	30.5	33.12	35.02	36.34	38.58	39.96
504	19.5	23.85	25.74	28.16	26.9	29.18	32.44	35.5	32.16	30.22	28.53
505	20.6	27.35	27.3	29.68	32.2	34.91	38.23	42.81	43.1	42.77	43.44
558	22.45	24.7	27.14	29.02	28.52	30.05	31.94	34.02	34.73	34.9	35.86
626	20.44	25.1	27.21	27.17	29.01	31.3	32.32	34.25	36.3	39.13	39.1
629	21.84	24.39	25.8	27.28	28.83	28.45	30.07	30.8	32.23	32.7	32.5
630	21.48	23.99	25.85	27.18	27.42	28.68	30.85	31.6	33.78	35.3	34.9
785	20.3	24	25.8	28.3	29	30.9	30.6	32.6	33.8	35.9	35.8
849	20.8	22.2	23.4	24.2	25.2	27.2	28.2	29.4	31.2	32.7	28.7
831	21.1	25.5	25.8	26.8	29.9	34.1	35.6	38.1	40	42	43.2
872	19	20.8	22.3	22.9	24.3	25.2	24.9	25	25.4	26.7	27
832_2	20.8	24	24.3	26.3	27.9	29.6	31.5	31.6	33.4	34.6	36.2
842	21.1	23.6	24.8	27.5	29.1	29.9	32.1	33.8	34.4	36.2	35.9

Supplemental Table S3.2, continued

					UCE5 +/-						
841	22.2	23.99	25.8	28.7	30.38	32.87	35.27	37.32	38.73	40	41.49
489	21.2	22.8	25.16	25.35	27.15	31.08	33.63	35.66	37.65	38.02	39.42
490	20.2	22.45	25.45	28.28	29.05	31.54	34.96	37.08	38.47	40.6	41.03
559	19.84	22.27	23.84	25.7	25.37	27	29.15	30.15	31.3	31.6	32.45
623	17.4	25.5	24.62	23.06	24.85	26.47	27.85	29.06	30.1	32.43	34
665	19	23.23	26.25	28.96	30.29	31.6	34.15	36.2	39.4	40.4	39.5
667	21.48	24.95	26.25	27.93	29.05	30.6	32.14	33.8	36.1	37.4	37.5
788	23.7	27.3	29.8	32.1	34.7	37.8	40	43.2	43.9	46.2	49.1
784	19	22	24.3	28.1	29	32	30	30.3	32.6	34.5	35.3
786	21.9	25	26.6	28.2	29.9	31.5	31.5	33.2	33.4	35.9	36.8
830	20.5	24.2	24.5	26.2	27.6	30.6	32.6	35.7	37.8	39.8	40.6
832	20.7	23.5	25.6	24.1	26.9	30.2	32.4	34.5	36.1	37.7	39.2
870	22.1	24.4	24.7	25.4	25.8	27.3	29.2	29	27.9	28.4	29.7
871	22	24.4	24.9	26.8	29.2	31.1	34.2	35.5	36.5	37	38.1
873	22.7	25.3	26.9	28.3	29.6	30.3	31	32.3	34	35.8	37.4
831	22.5	24.3	25.9	28.4	30	31	34	33.8	36.3	38.8	41.6
					UCE5 -/-						
446	18.03	20	21.1	22.5	24.3	25.8	26.4	26.3	27.2	29.2	30.46
449	21.45	24.6	25.9	28.02	31	33.23	35.09	36.01	38.3	40.63	41.56
839	19.4	21.2	22.6	24.46	25.21	26.96	28.29	29.88	31.15	32.36	31.3
507	21	24.9	25.6	27.19	28.24	29.7	30.8	32.6	33.4	32.5	34.22
506	14.6	16.5	17.7	18.57	20.18	21	21.75	21.8	20.55	20.49	21.5
557	20.83	23.92	25.5	27.06	27.84	28.49	30.13	30.59	31.9	32.3	32.56
624	18.16	20	21.59	21.05	22.08	23.76	24.95	25.66	27.2	28.85	31.1
625	18.18	23.4	23.9	24.24	26.44	28.68	30.34	31.15	33.2	35.84	38.7
622	17	20.21	21.3	21.62	22.7	24.61	25.02	26.44	27.6	28.9	30
633	21.09	24.22	26.39	27.71	30.4	32.98	34.34	35.8	39.26	41.9	42
632	21.28	24.05	23.49	25.28	26.67	28.96	31.14	31.8	34.48	36.4	35.9
666	16.56	19.57	21.32	23.07	23.9	25.3	27.1	27.6	29.5	30.9	31.9
783	19.8	23.2	26	28.8	29.7	32	30.5	31.4	33.2	34.6	35
787	21	24.8	26.5	28.4	30.4	32.6	32.8	34.3	35.4	37.4	37.5
828	20.6	21.2	22	22.4	23.2	23.8	24.9	26.5	27.5	28.9	29.8
829	20.4	21.2	21.4	22.4	23.2	24.8	25.9	27.2	28.2	29	28.5
834	21	23.3	24.2	24.5	26.7	30.3	32.8	35.5	36.4	39.8	40.6
833	20.5	24	23.4	24.1	25.5	27.8	29.7	30.7	32.1	33.7	35.4
869	22.4	23.8	26.2	28.5	30.4	32.4	34.9	36	37.7	37.8	37.8
900	21.5	25.8	27.7	30.2	31.3	31.8	34.3	34.5	35	36.5	38.1

Supplemental Table S3.2, continued

833_2	20.7	23.5	24.7	26.7	28.7	29	31	30.9	31.9	33.6	35.8
841	21.5	22.5	23.1	23.6	24.6	24.9	26.7	26.8	28.1	27.5	26.5
840_2	20.9	21.5	22.9	24.7	24.7	24.4	24.8	25.7	26.6	26.4	25.2

Supplemental Table S3.3 Body weight phenotypes resulting from enhancer deletions (literature search).

Enhancer deletion	Size of deletion	target gene	distance to target gene	BW pheno type?	Increase/ Decrease?	Ref.
EE enhancer	1 kb	Hoxc8	3 kb upstream	yes	decrease	Juan and Ruddle (2003)[123]
Pomc enhancer nPE1	579 bp	Pomc	10kb upstream	yes	increase	Lam et al. (2015)[124]
Pomc enhancer nPE2	172 bp	Pomc	10 kb upstream	no		Lam et al. (2015)[124]
H19 enhancer	6.2 kb	H19/Igf2 impriting	10 kb	yes	decrease	Leighton et al. (1995)[125]
CAD 70kb association	70 kb	CDKN2A/ B	50 kb	yes	increase	Visel et al. (2010)[126]
MFCS1 enhancer	1167 bp	Shh	1 Mb	yes	decrease	Sagai et al. (2005)[127]
mm771 heart enhancer	332 bp	Myh7	2.5 kb	yes	decrease (female)	Dickel et al. (2016)[128]
M280	9.7 kb	Tmem53, Dmap1	>100 kb	yes	decrease	Nolte et al. (2014)[27]
M1442	7.7 kb	Kifap3, Nme7	500 kb	no	N/A	Nolte et al. (2014)[27]
hs_119	2.1 kb	Arx	122 kb	yes	decrease	Dickel et al. (2018)[28]
hs_121 (aka UCE_467)	2.8 kb	Arx	24 kb	no	N/A	Dickel et al. (2018)[28]
hs123		Arx	500 kb	no	N/A	Dickel et al. (2018)[28]

Supplemental Table S3.3, continued

SBE1	525 bp	Shh	6.6 kb	yes	decrease, severe	Jeong et al. (2011)[129]
RSCE enhancer	446 bp	Sox9	1 Mb	no	N/A	Mochizuki et al. (2018)[130]
Myc super-enhancer	538 kb	Myc	2-540 kb	no	N/A	Dave et al. (2017)[131]
Myc-335	1.7 kb	Myc	335 kb	no	N/A	Sur et al. (2012)[132]
IG-DMR CS1	438 bp	Dlk1-Dio3	~100 kb downstream	no	N/A	Saito et al. (2017)[133]
IG-DMR CS2	292 bp	Dlk1-Dio3	~100 kb downstream	no	N/A	Saito et al. (2017)[133]
IG-DMR CS3	303 bp	Dlk1-Dio3	~100 kb downstream	no	N/A	Saito et al. (2017)[133]
I56ii Dlx enhancers	2 kb	Dlx	intergenic (<1kb)	no	N/A	Darbandi et al. (2016)[134]
Ebeta enhancer	560 bp	TCR	< 10kb	no	N/A	Bouvier et al. (1996)[135]
BENC enhancer	200 kb?	Myc	1.7 Mb	no	N/A	Bahr et al. (2018)[136]
D2 enhancer	2.1 kb	Tnfsf11	23kb	no	N/A	Onal et al. (2016)[137]
mm77 heart enhancer	2.5 kb	Myl2	7 kb	no	N/A	Onal et al. (2016)[137]
D6 enhancer	1 kb	Tnfsf11	83 kb	no	N/A	Onal et al. (2016)[138]
T1 enhancer	7.4 kb	Tnfsf11	123 kb	no	N/A	Onal et al. (2016)[138]
Renin enhancer (RE)	350 bp	Renin	2.7 kb	no	N/A	Adams et al. (2006)[139]

Supplemental Table S3.3, continued

UCE_248	not avail.	DMRT1/2/3	>100 kb	no	N/A	Ahituv et al. (2007)[26]
UCE_329	not avail.	Pax6/Wt1	>200 kb	no	N/A	Ahituv et al. (2007)[26]
UCE_467	not avail.	Arx	24 kb	no	N/A	Ahituv et al. (2007)[26]
UCE_482	not avail.	Sox3	350 kb	no	N/A	Ahituv et al. (2007)[26]

CHAPTER4: A PROMOTER INTERACTION MAP FOR CARDIOVASCULAR DISEASE GENETICS

4.1 Abstract[2]

Over 500 genetic loci have been associated with risk of cardiovascular diseases (CVDs), however most loci are located in gene-distal non-coding regions and their target genes are not known. Here, we generated high-resolution promoter capture Hi-C (PCHi-C) maps in human induced pluripotent stem cells (iPSCs) and iPSC-derived cardiomyocytes (CMs) to provide a resource for identifying and prioritizing the functional targets of CVD associations. We validate these maps by demonstrating that promoters preferentially contact distal sequences enriched for tissue-specific transcription factor motifs and are enriched for chromatin marks that correlate with dynamic changes in gene expression. Using the CM PCHi-C map, we linked 1,999 CVD-associated SNPs to 347 target genes. Remarkably, more than 90% of SNP-target gene interactions did not involve the nearest gene, while 40% of SNPs interacted with at least two genes, demonstrating the importance of considering long-range chromatin interactions when interpreting functional targets of disease loci.

4.2 Introduction

A major goal in human genetics research is to understand genetic contributions to complex diseases, specifically the molecular mechanisms by which common DNA variants impact disease etiology. Most genome-wide association studies (GWAS) implicate non-coding variants that are far from genes, complicating interpretation of their mode of action and correct identification of the

[2] Reproduced with permission from: Montefiori, L. E. *et al.* A promoter interaction map for cardiovascular disease genetics. *Elife* **7,** 1–35 (2018).

target gene[140]. Mounting evidence suggests that disease variants disrupt the function of *cis*-acting regulatory elements, such as enhancers, which in turn affects expression of the specific gene or genes that are functional targets of these elements[74,82,110,141,142]. However, because *cis*-acting regulatory elements can be located kilobases (kb) away from their target gene(s), identifying the true functional targets of regulatory elements remains challenging[74].

Chromosome conformation capture techniques such as Hi-C[36] enable the genome-wide mapping of long-range chromatin contacts and therefore represent a promising strategy to identify distal gene targets of disease-associated genetic variants. Recently, Hi-C maps have been generated in numerous human cell types including embryonic stem cells and early embryonic lineages[40,41], immune cells[143], fibroblasts[144] and other primary tissue types[145]. However, despite the increasing abundance of Hi-C maps, most datasets are of limited resolution (>40 kb) and do not precisely identify the genomic regions in contact with gene promoters.

More recently, promoter capture Hi-C (PCHi-C) was developed which greatly increases the power to detect interactions involving promoter sequences[49,50]. PCHi-C in different cell types identified thousands of enhancer-promoter contacts and revealed extensive differences in promoter architecture between cell types and throughout differentiation[49,50,146–149]. These studies collectively demonstrated that genome architecture reflects cell identity, suggesting that disease-relevant cell types are critical for successful interrogation of the gene regulatory mechanisms of disease loci.

In support of this notion, several recent studies utilized high-resolution promoter interaction maps to identify tissue-specific target genes of GWAS associations. Javierre *et al*. generated promoter capture Hi-C data in 17 primary human blood cell types and identified 2,604 potentially causal genes for immune- and blood-related disorders, including many genes with unannotated roles in those diseases[146]. Similarly, Mumbach *et al*. interrogated GWAS SNPs

associated with autoimmune diseases using HiChIP where they identified ~10,000 promoter-enhancer interactions that linked several hundred SNPs to target genes, most of which were not the nearest gene[78]. Importantly, both studies reported cell-type specificity of SNP-target gene interactions.

Cardiovascular diseases, including cardiac arrhythmia, heart failure, and myocardial infarction, continue to be the leading cause of death world-wide. Over 50 GWAS have been conducted for these specific cardiovascular phenotypes alone, with more than 500 loci implicated in cardiovascular disease risk (NHGRI GWAS catalog, https://www.ebi.ac.uk/gwas/), most of which map to non-coding genomic regions. To begin to dissect the molecular mechanisms by which genetic variants contribute to CVD risk, a comprehensive gene regulatory map of human cardiac cells is required. Here, we present high resolution promoter interaction maps of human iPSCs and iPSC-derived cardiomyocytes (CMs). Using PCHi-C, we identified hundreds of thousands of promoter interactions in each cell type. We demonstrate the physiological relevance of these datasets by functionally interrogating the relationship between gene expression and long-range promoter interactions, and demonstrate the utility of long-range chromatin interaction data to resolve the functional targets of disease-associated loci.

4.3 Results

4.3.1 iPSC-derived cardiomyocytes provide an effective model to study the architecture of CVD genetics

We used iPSC-derived CMs[150] as a model to study cardiovascular gene regulation and disease genetics. The CMs generated in this study were 86-94% pure based on cardiac Troponin T protein expression and exhibited spontaneous, uniform beating (Supplemental Figure S4.1A).

To demonstrate that iPSCs and CMs recapitulate transcriptional and epigenetic profiles of matched primary cells, we conducted RNA-seq and ChIP-seq for the active enhancer mark H3K27ac in both cell types and compared these data with similar cell types from the Epigenome Roadmap Project[13]. RNA-seq profiles of iPSCs clustered tightly with H1 embryonic stem cells, whereas CMs clustered with both left ventricle (LV) and fetal heart (FH) profiles (Supplemental Figure S4.1B). Furthermore, we observed that matched cell types exhibited three-fold greater overlap in the number of promoter-distal H3K27ac ChIP-seq peaks than non-matched cell types (Supplemental Figure S4.1C,D), indicating that both iPSCs and CMs recapitulate tissue-specific epigenetic states of human stem cells and primary cardiomyocytes, respectively.

To further validate our system, we analyzed differentially expressed genes between iPSCs and CMs. Among the top 10% of over-expressed genes in CMs were genes directly related to cardiac function including essential cardiac transcription factors (*GATA4*, *MEIS1*, *TBX5*, and *TBX20*) and differentiation products (*TNNT2*, *MYH7B*, *MYL7*, *ACTN2*, *NPPA*, *HCN4*, and *RYR2*) (fold-change > 1.5, $P_{adj} < 0.05$, Supplemental Figure S4.2A-C). Gene Ontology (GO) enrichment analysis for genes over-expressed in CMs relative to iPSCs further confirmed the cardiac-specific phenotypes of these cells with top terms relating to the development of the cardiac conduction system and cardiac muscle cell contraction (Supplemental Figure S4.2D).

4.3.2 Promoter-capture Hi-C identifies distal regulatory elements in iPSCs and CMs

To comprehensively map long-range regulatory elements in iPSCs and CMs, we performed *in-situ* Hi-C[143] in triplicate iPSC-CM differentiations; importantly, we used the 4-cutter restriction enzyme MboI which generates ligation fragments with an average size of 422 bp, enabling enhancer-level resolution of promoter contacts. We enriched iPSC and CM *in situ* Hi-C libraries for promoter interactions through hybridization with a set of 77,476 biotinylated RNA probes

("baits") targeting 22,600 human RefSeq protein-coding promoters (see section 4.5 Methods) and sequenced each library to an average depth of ~413 million (M) paired-end reads. After removing duplicates and read-pairs that did not map to a bait, we obtained an average of 31M and 41M read-pairs per replicate for iPSC and CM, respectively. We used CHiCAGO[151], a computational pipeline which accounts for bias from the sequence capture, to identify significant interactions and further filtered for those significant in at least two out of three replicates (see section 4.5 Methods). Finally, we exclusively focused on interactions that were separated by a distance of at least 10 kb. This criterion addresses the high frequency of close-proximity ligation evets in Hi-C data, which are difficult to distinguish as random Brownian contacts or functional chromatin interactions[151]. In total, we identified 350,062 promoter interactions in iPSCs and 401,098 in CMs. A large proportion (~55%) of interactions were shared between the two cell types, indicating that even at high resolution many long-range interactions are stable across cell types (Figure 4.1A). Approximately 20% of all interactions were between two promoters, demonstrating the high connectivity between genes and supporting the recently suggested role of promoters acting as regulatory inputs for distal genes[152,153] (Figure 4.1B). Most interactions were promoter-distal, with a median of ~170 kb between the promoter and the distal-interacting region (Figure 4.1C).

Figure 4.1 General features of promoter interactions. (A) Venn diagram displaying the number of cell type-specific and shared promoter interactions in each cell type.

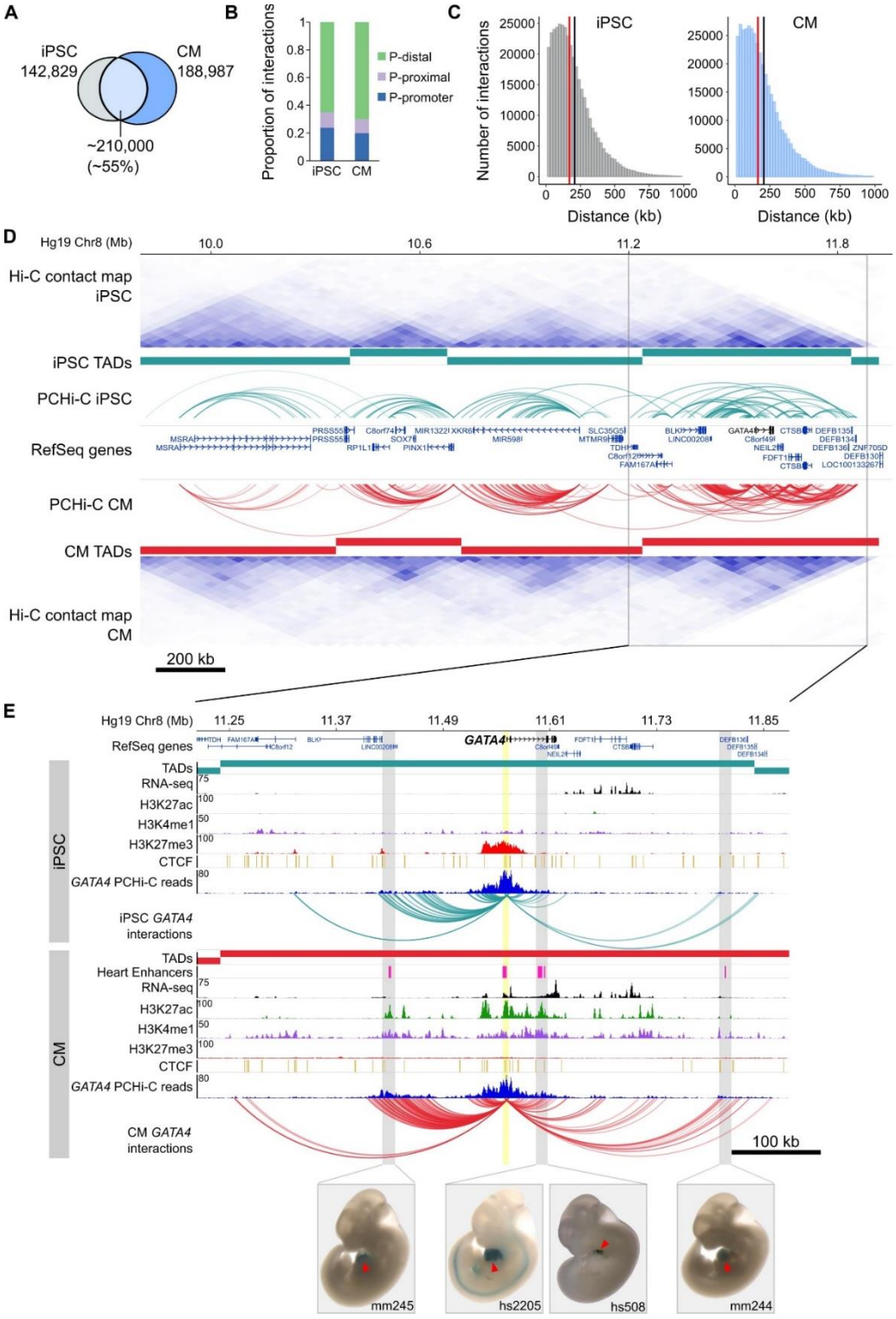


Figure 4.1, continued. (B) Proportion of interactions in each distance category: promoter (P)-promoter (both interacting ends overlap a transcription start site (TSS)); P-proximal (non-promoter end overlaps captured region but not the TSS); P-distal (non-promoter end is outside of captured region). Note that all promoter interactions are separated by at least 10 kb. **(C)** Distribution of the distances spanning each interaction in iPSCs and CMs. The red line depicts the median (170 kb in iPSCs, 164 kb in CMs); the black line depicts the mean (208 kb in iPSCs, 206 kb in CMs). **(D)** A ~2 Mb region of chromosome 8 encompassing the *GATA4* gene is shown along with pre-capture (whole genome) Hi-C interaction maps at 40 kb resolution for iPSCs (top) and CMs (bottom). TADs called with TopDom are shown as colored bars (median TAD size = 640 kb in both cell types, mean TAD size = 742 kb in iPSCs and 743 kb in CMs) and significant PCHi-C interactions as colored arcs. **(E)** Zoomed-in view of the *GATA4* locus (promoter highlighted in yellow) in iPSCs (top) and CMs (bottom) along with corresponding RNA-seq data generated as part of this study, and ChIP-seq data for H3K27ac, H3K4me1, H3K27me3 and CTCF from the Epigenome Roadmap Project/ENCODE (H1 and left ventricle for iPSC and CM, respectively). Filtered *GATA4* read counts used by CHiCAGO are displayed in blue with the corresponding significant interactions shown as arcs. For clarity, only *GATA4* interactions are shown. Gray highlighted regions show interactions overlapping *in vivo* validated heart enhancers (pink boxes), with representative E11.5 embryos for each enhancer element[111]. Red arrowhead points to the heart.

To compare the PCHi-C maps with known features of genome organization, we sequenced our pre-capture Hi-C libraries to an average depth of 665M reads per cell type and identified topologically associating domains (TADs) with TopDom (see section 4.5 Methods). TADs are organizational units of chromosomes defined by <1 megabase (Mb) genomic blocks that exhibit high self-interacting frequencies with a very low interaction frequency across TAD boundaries[41,42]. Notably, this organization is thought to constrain the activity of *cis*-regulatory elements to target genes with the same TAD, as disruption of TAD boundaries has been shown to lead to aberrant activation of genes in neighboring TADs[42,45–48]. We found that the majority of PCHi-C interactions occurred within TADs (73% and 77% in iPSCs and CMs, respectively; Figure 4.1D and Supplemental Figure S4.3A). TAD-crossing interactions ("inter-TAD") contained proportionally more promoter-promoter interactions than intra-TAD interactions, and were more likely to overlap promoter-distal CTCF sites; however, they were similarly enriched for looping to distal H3K27ac sites, a mark of active chromatin (Supplemental Figure S4.3B-D). Inter-TAD interactions had slightly lower CHiCAGO scores, reflecting a lower number of reads supporting these interactions, and spanned greater genomic distances than intra-TAD interactions (Supplemental Figure S4.3E,F). Additionally, promoters with inter-TAD interactions were preferentially located close to TAD boundaries (Supplemental Figure S4.3G) and had higher expression levels compared to promoters with intra-TAD interactions, particularly in CMs (Supplemental Figure S4.3H). These observations are consistent with previous studies which demonstrated that highly expressed genes, specifically housekeeping genes, are enriched at TAD boundaries[41].

To illustrate the utility of high-resolution PCHi-C interaction maps, we highlight the *GATA4* locus in Figure 4.1D,E). *GATA4* is a master regulator of heart development[154,155] and the *GATA4* gene is located in a TAD structure that is relatively stable between iPSCs and CMs (Figure

4.1D). However, PCHi-C identified increased interaction frequencies between the *GATA4* promoter and several H3K27ac-marked regions, including four *in vivo* validated heart enhancers from the Vista enhancer browser[111], specifically in CMs and coincident with strong up-regulation of *GATA4* (Supplemental Figure S4.2C). Although TAD-based analyses help define a gene's *cis*-regulatory landscape, high resolution promoter interaction data provides the resolution necessary to precisely map enhancer-promoter interactions in the context of cellular differentiation.

To validate the CM interaction map as a resource for cardiovascular disease genetics we next extensively characterized several important aspects of genetic architecture in CMs. We compared CMs with iPSCs in each analysis as a measure of cell-type specificity. These analyses serve as benchmarks that build on established features of genome organization and aid interpretations of the roles that long range interactions play in gene regulation.

4.3.3 Promoter interactions are enriched for tissue-specific transcription factor motifs

Distal enhancers activate target genes through DNA looping, a mechanism that enables distally bound transcription factors to contact the transcription machinery of target promoters[156–158]. To assess whether this feature of gene regulation was reflected in the iPSC and CM interactions, we conducted motif analysis using HOMER[90] on the set of promoter-distal interacting sequences in each cell type. We initially focused on interactions for genes differentially expressed between iPSCs and CMs (fold-change > 1.5, $P_{adj} < 0.05$). We identified CTCF as the most enriched motif in each case (Figure 4.2A,B), consistent with the known role of this factor in mediating long-range genomic interactions[159–161]. Among the other top motifs, we identified the pluripotency factor motifs OCT4-SOX2-TCF-NANOG (OSN) and SOX2 as preferentially enriched in distal sequences looping to genes over-expressed in iPSCs (Figure 4.2A,C), whereas top motifs in distal sequences looping to genes over-expressed in CMs included TBX20, ESRRB and MEIS1 (Figure

4.2B,C). TBX20 and MEIS1 transcription factors are important regulators of heart development and function[162–164] and ESRRB was previously identified as a potential binding partner of TBX20 in adult mouse cardiomyocytes[165]. We also observed that distal interactions unique to either iPSCs or CMs were similarly enriched for tissue-specific transcription factor motifs (Figure 4.2D). In line with a recent report that AP-1 contributes to dynamic loop formation during macrophage development[166], both iPSC- and CM-specific interactions were enriched for AP-1 motifs (Figure 4.2D), suggesting that AP-1 transcription factors may represent a previously unrecognized genome organizing complex.

Figure 4.2 Transcription factor motif enrichment in distal interacting regions. (A,B) Selected transcription factor (TF) motifs identified using HOMER in the promoter-distal interacting sequences for all over-expressed genes in **(A)** iPSCs and **(B)** CMs (fold change > 1.5, $P_{adj} < 0.05$). "% sites" refers to the percent of distal interactions overlapping the motif; rank is based on P-value significance. **(C)** To compare motif ranks across gene sets, the inverse of the rank is plotted for selected motifs identified in distal interactions from over- or under-expressed genes in both iPSCs and CMs. **(D)** The top 50 motifs identified in cell type-specific interactions. *OSN*, OCT4-SOX2-TCF-NANOG motif.

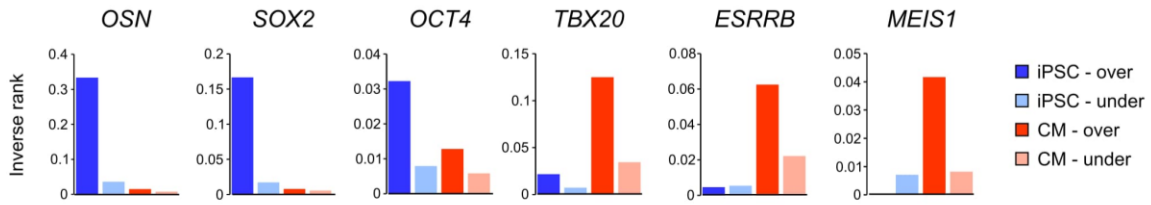
A

Motif	Logo	% sites	Rank	P-value
CTCF		4.85%	1	1 x 10^{-107}
OSN		6.51%	3	1 x 10^{-20}
SOX2		22.46%	6	1 x 10^{-14}
OCT4		13.30%	31	1 x 10^{-8}

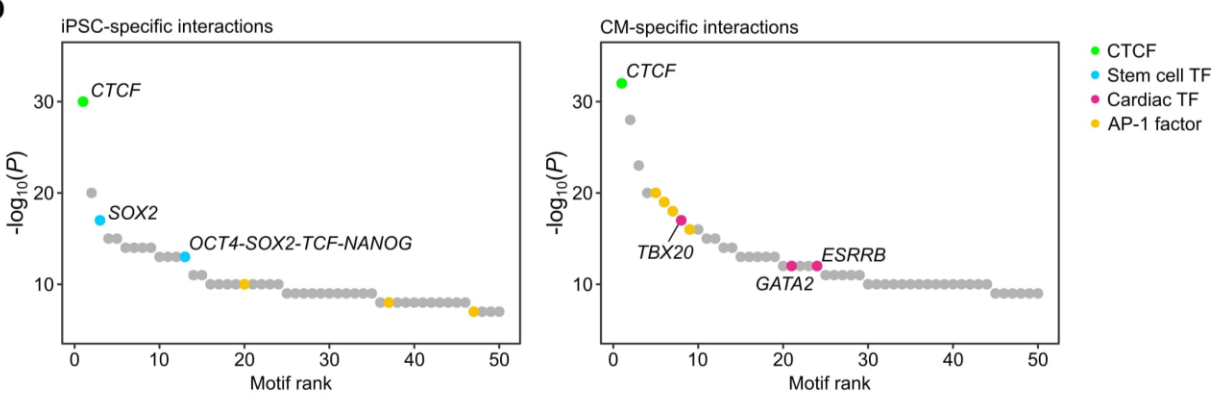
B

Motif	Logo	% sites	Rank	P-value
CTCF		4.59%	1	1 x 10^{-100}
TBX20		6.46%	8	1 x 10^{-20}
ESRRB		17.7%	16	1 x 10^{-18}
MEIS1		41.06%	24	1 x 10^{-16}

C



D



4.3.4 Long-range promoter interactions are enriched for active cis-regulatory elements and correspond to gene expression dynamics

Functionally active *cis*-regulatory elements are characterized by the presence of specific histone modifications; active enhancers are generally associated with H3K4me1 and H3K27ac[4,167], whereas inactive (e.g. poised or silenced) elements are often associated with H3K27me3[168,169]. In support of the gene-regulatory function of long-range interactions, we found that the promoter-distal MboI fragments involved in significant promoter interactions were enriched for these three histone modifications in both iPSCs and CMs (Figure 4.3A-C). When promoters were grouped by expression level, we observed that this enrichment increased with increasing expression for H3K27ac and H3K4me1, and decreased with increasing expression for H3K27me3, consistent with an additive nature of enhancer-promoter interactions[49,146], and validating that PCHi-C enriches for likely functional long-range chromatin contacts.

A strong correlation (Pearson correlation coefficient $r > 0.7$) between the degree of histone modifications and gene expression was first reported nearly ten years ago[170], however that analysis only considered histone modifications within 2 kb of promoters. To understand whether this relationship extends beyond promoter-proximal regions, we correlated the number of histone ChIP-seq peaks within 300 kb of promoters with the promoter's expression level (Supplemental Figure S4.4A,B). H3K27ac and H3K4me1 both positively correlated with expression level (Spearman's ρ = 0.22 and 0.16, respectively in iPSC and ρ = 0.23 and 0.24, respectively in CMs, $P < 2.2^{-16}$); in contrast, H3K27me3 negatively correlated with expression level in CMs (Spearman's ρ = -0.20, $P < 2.2^{-16}$), however this relationship was not present in iPSCs (Spearman's ρ = 0.02, P = 0.06). Although moderate, these correlations could partially explain why higher expressed genes show stronger enrichment for promoter interactions overlapping histone peaks

when using a genome-wide background model (see section 4.5 Methods), and lends support to the notion that active genes are located in generally active genomic environments[171,172].

We next investigated the relationship between cell type-specific interactions and enrichment for tissue-specific CTCF, H3K27ac, and H3K27me3 marks, hypothesizing that interactions unique to iPSCs or CMs would be most enriched for tissue-specific chromatin features. Indeed, we observed that cell type-specific interactions preferentially involved H3K27ac peaks from the matched cell type, and were either not enriched (iPSC) or depleted (CM) for H3K27ac marks that were specific to the non-matched cell type (Figure 4.3E, middle panel). However, the strongest enrichment was for cell type-specific interactions to overlap chromatin features that were present in both cell types (Figure 4.3E). Additionally, interactions that were shared between iPSCs and CMs were most enriched for shared chromatin features. These results suggest that all interactions, whether shared or unique to one cell type, preferentially contact regulatory regions that are active in both cell types, whereas cell type-specific interactions are not likely to occur in regions specifically marked in the non-matched cell type.

An example of a gene that encompasses these observations is the atrial natriuretic peptide gene *NPPA* (Figure 4.3F) which is specifically expressed in cells of the heart atrium and is upregulated in CMs (Supplemental Figure S4.2). *NPPA* makes numerous cell type-specific interactions to a distal region that is only marked with active chromatin (H3K27ac and H3K4me1) in CMs; furthermore, functional characterization showed that this region corresponds to an *in vivo* enhancer recapitulating *NPPA*'s endogenous expression in the developing heart[111]. Taken together, these results illuminate the complex relationship between long-range promoter interactions and gene regulation and provide evidence that promoter architecture reflects cell type-specific gene expression.

Figure 4.3 Enrichment of promoter interactions to distal regulatory features. (A,B) Proportion of promoter-distal interactions overlapping a histone ChIP-seq peak compared to random control MboI fragments (see section 4.5 Methods). iPSC interactions were overlapped with H1 ESC ChIP-seq data; CM interactions were overlapped with left ventricle ChIP-seq data from the Epigenome Roadmap Project (Supplemental File S4.10).

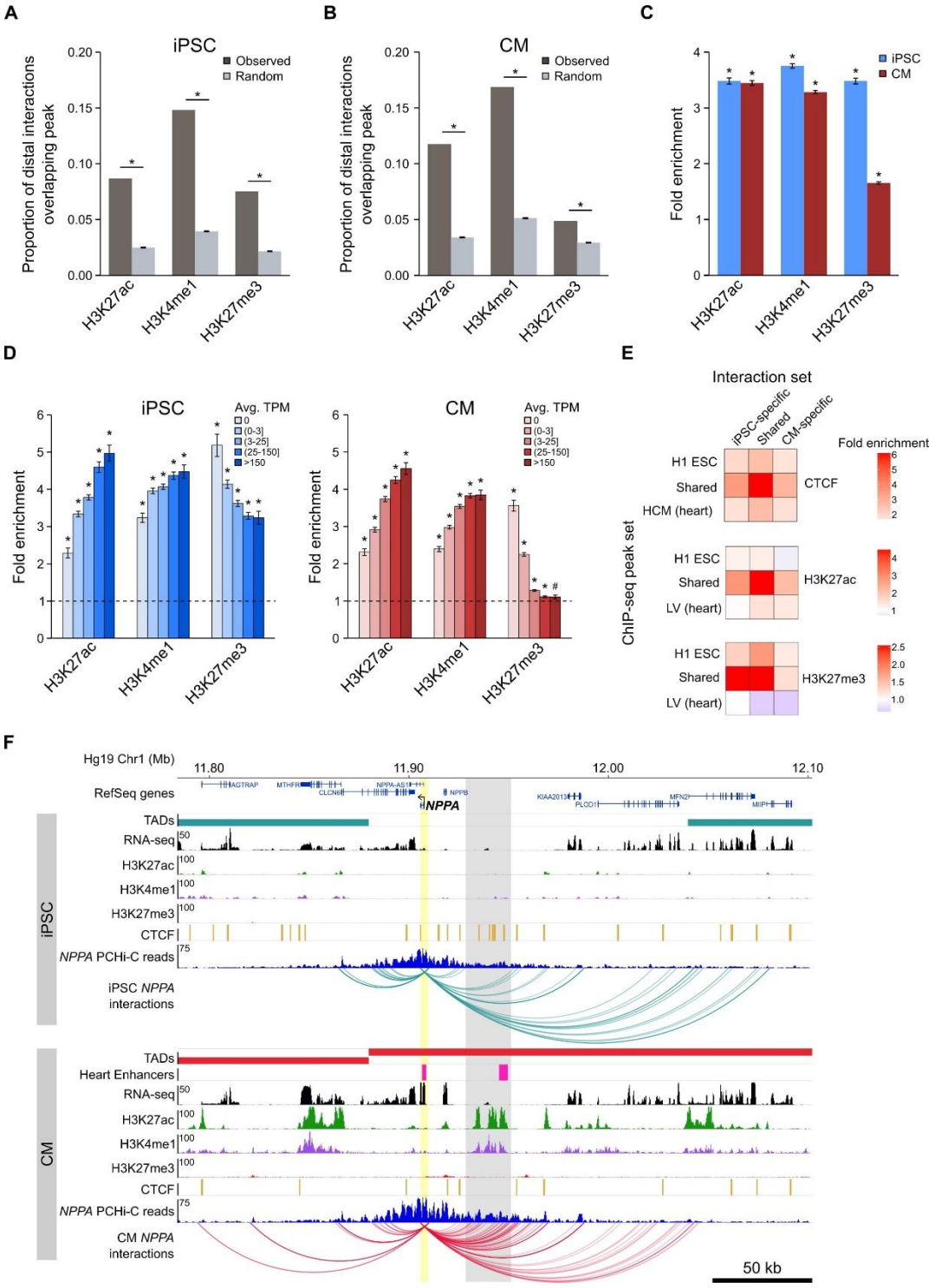


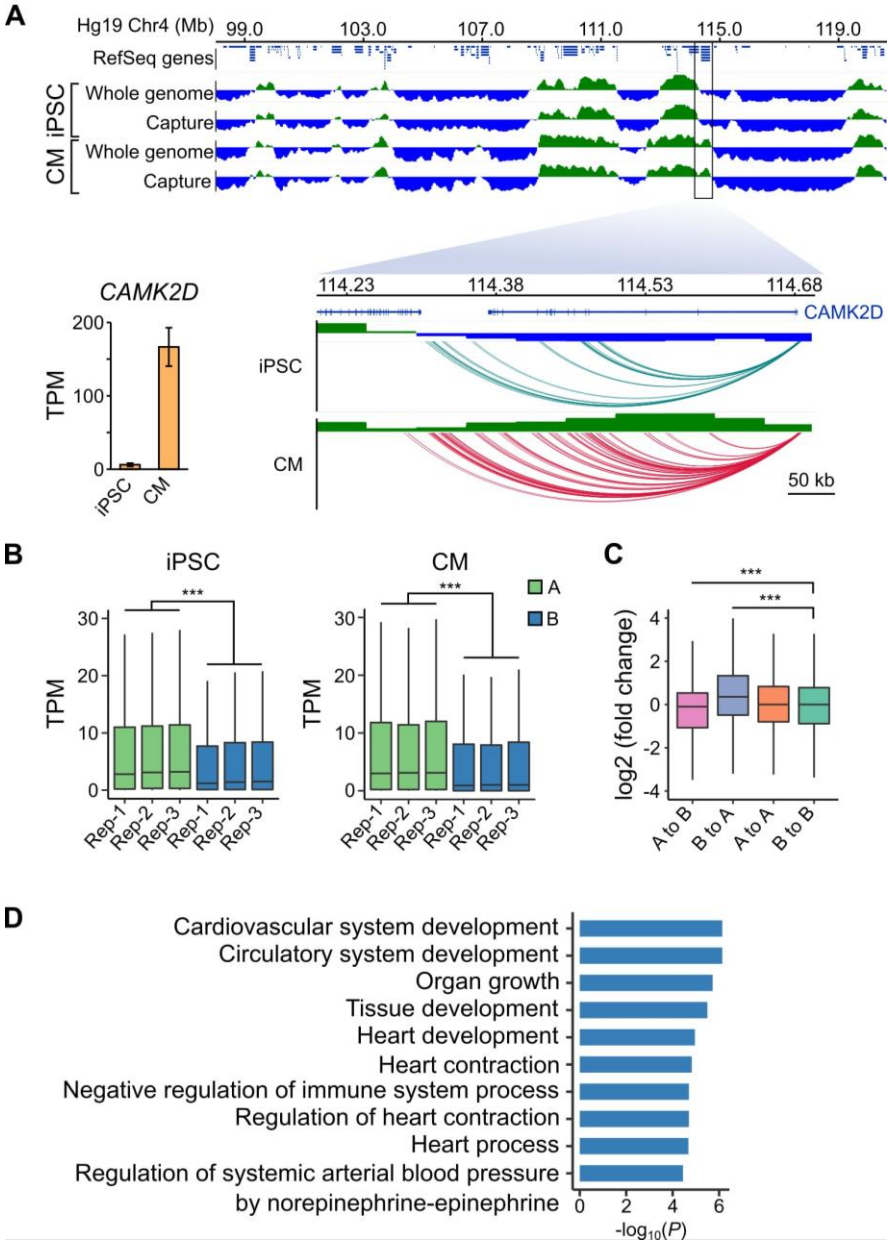
Figure 4.3, continued. (C) Fold enrichment of the data presented in **(A)** and **(B)**. **(D)** Fold enrichment of promoter-distal interactions based on the expression level of the promoter. Promoters were grouped into 5 bins according to their average TPM values. Dashed line indicates no enrichment. **(E)** Fold enrichment of cell type-specific and shared interactions (columns) to tissue-specific and shared chromatin features (rows). **(F)** Example of the *NPPA* gene in iPSCs (top) and CMs (bottom). Gray box highlights CM-specific interactions to CM-specific chromatin marks and an *in vivo* heart enhancer[111]. For clarity, only interactions for *NPPA* are shown. $*P < 0.00001$, $^{\#}P = 0.0017$, Z-test.

4.3.5 Dynamic changes in genomic compartmentalization involve a subset of cardiac-specific genes

As a final benchmark of our datasets, we analyzed large-scale differences in genome organization between iPSCs and CMs. The first Hi-C studies revealed that the genome is organized in two major compartments, A and B, that correspond to open and closed regions of chromosomes, respectively[36,143]. Although most compartments are stable across different cell types, some compartments switch states in a cell type-specific manner which may reflect important gene regulatory changes[40]. To assess whether capture Hi-C data, which is more cost-effective for capturing promoter-centered interactions, is able to identify A/B compartments, we compared our capture Hi-C data with pre-capture, genome-wide Hi-C libraries. A/B compartments identified using HOMER[90] were remarkably similar in the whole-genome and PCHi-C datasets (97% correspondence, Figure 4.4A, top panel, and Supplemental Figures S4.5 and S4.6), demonstrating that PCHi-C data contains sufficient information to identify broadly active and inactive regions of the genome. As an example, we highlight a 10 Mb region on chromosome 4 containing the *CAMK2D* gene locus (Figure 4.4A). Compartments were relatively stable across this region in iPSCs and CMs, however the *CAMK2D* gene itself was located in a dynamic compartment that switched from inactive in iPSCs to active in CMs. Correspondingly, this gene was highly upregulated during differentiation to CMs (Figure 4.4A, inset).

We observed this effect on a global level, as genes located in A compartments were expressed at significantly higher levels than genes located in the B compartments in both iPSCs and CMs (Figure 4.4B). Additionally, genes that switched A/B compartments between cell types were correspondingly up- or down-regulated (Figure 4.4C). GO analysis of the 1,008 genes that switched from B to A compartments during iPSC-CM differentiation revealed enrichment for terms such as "cardiovascular system development" and "heart contraction" (Figure 4.4D, Supplemental File S4.5). Importantly, these genes were identified based solely on their location in a dynamic genomic compartment and not from gene expression data. GO analysis for genes that switched from A to B compartments during iPSC-CM differentiation related to non-cardiac processes, such as skin development, epithelial cell differentiation and sex determination (Supplemental Figure S4.7, Supplemental Files S4.5 and S4.6). These data show that PCHi-C accurately captured tissue-specific interactions and indicate that compartmentalization of genes in spatially regulated regions of the nucleus may be one mechanism to ensure tissue-specific gene expression[40]. In summary, our analyses demonstrated that CM promoter interactions recapitulate key features of cardiac gene regulation and function, validating the CM map as an important tool to investigate CVD genetics.

Figure 4.4. **A/B compartment switching corresponds to activation of tissue-specific genes. (A)** Top panel: 10 Mb region on chromosome 4 showing A (green) and B (blue) compartments based on the first principle component analysis calculated by HOMER [90] of the whole-genome Hi-C and capture Hi-C interaction data. Bottom panel: zoomed in on the *CAMK2D* locus; only capture Hi-C A/B compartments shown. Inset: expression level of *CAMK2D* in iPSCs and CMs across the three replicates. **(B)** Expression level (TPM) of genes located in the A (green) or B (blue) compartment in each replicate of iPSC (left) or CM (right). **(C)** Difference in expression level (log2 fold change relative to iPSCs) of genes switching compartments from iPSC to CM or remaining in stable compartments. **(D)** Gene Ontology analysis of biological processes associated with genes switching from B to A compartments during iPSC-CM differentiation. *** $P < 2.2 \times 10^{-16}$, Wilcoxon rank-sum test.



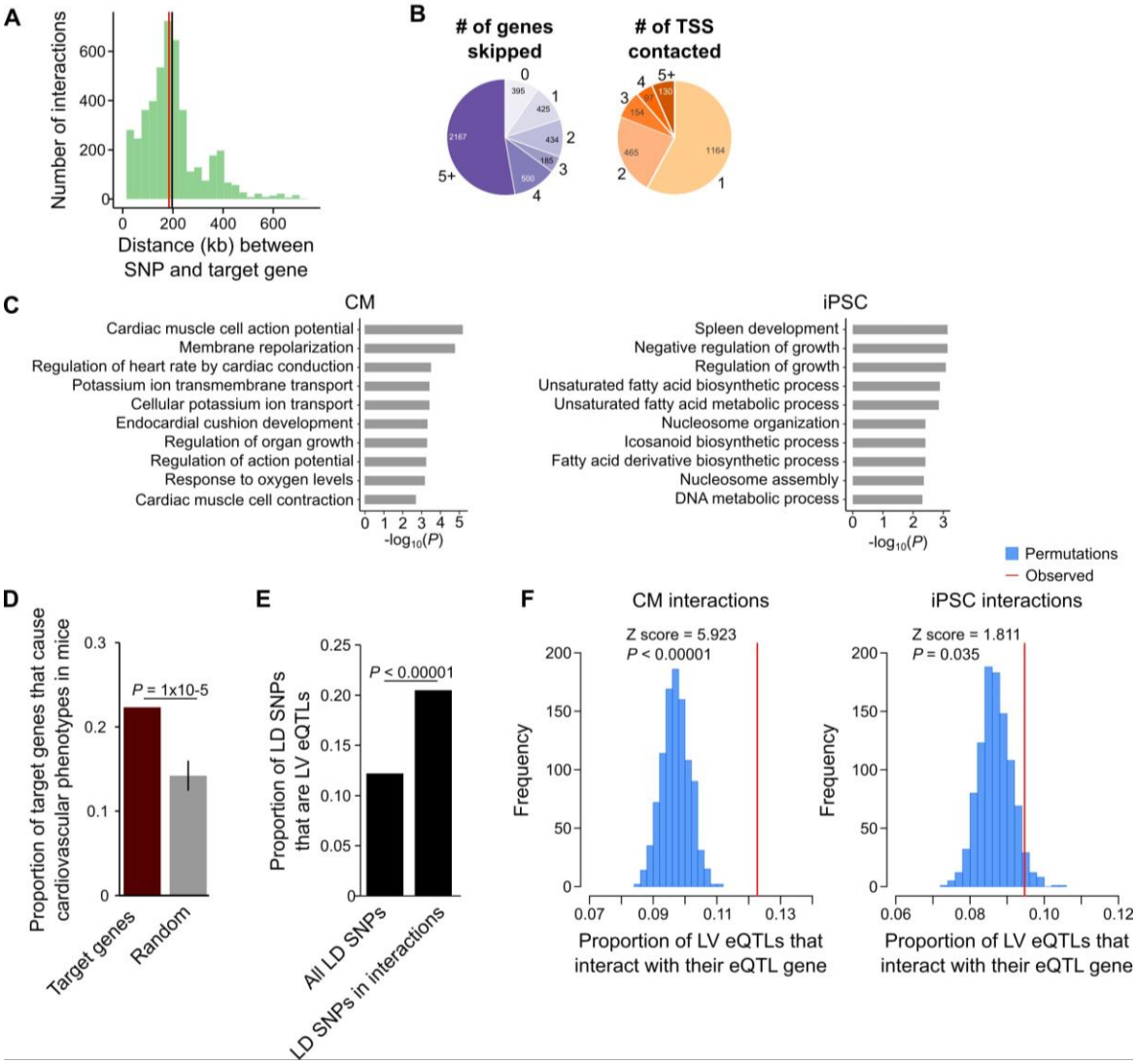
4.3.6 CM promoter interactions link GWAS SNPs to target genes

A particularly relevant application of high resolution promoter interaction maps is to guide post-GWAS studies by identifying the target genes of disease-associated variants. We employed this approach to link GWAS SNPs for several major cardiovascular diseases to their target gene(s) using the CM interaction map. We compiled 524 lead SNPs from the NHGRI-EBI database (https://www.ebi.ac.uk/gwas/) for three important classes of CVDs: cardiac arrhythmias, heart failure, and myocardial infarction (Supplemental File S4.7 and S4.8). Because of linkage disequilibrium (LD) patterns, the true causal SNP could be any SNP in high LD with the lead variant. Therefore, we expanded this set of SNPs to include all variants in high LD ($r^2 > 0.9$, within 50 kb of lead SNP), increasing the number of putatively causal variants to 10,475 (hereafter called LD SNPs). We found that 1,999 (19%) of the LD SNPs were located in promoter-distal MboI fragments that interacted with the promoters of 347 genes in CMs (Supplemental File S4.8), hereafter referred to as target genes. The majority (89%) of LD SNP-target gene pairs were located within the same TAD, with a median distance of 185kb between each SNP-target gene pair (Figure 4.5A). Importantly, 90.4% of SNP-target gene interactions skipped at least one gene promoter and 42% of SNPs interacted with at least two different promoters (Figure 4.5B).

To confirm that the CM PCHi-C interactions linked SNPs to CVD-relevant target genes, we performed GO analysis and found that target genes were highly and specifically enriched for biological processes related to cardiac function, such as membrane repolarization and cardiac conduction (Figure 4.5C, left panel and Supplemental File S4.5 and S4.6). As a control, we used iPSC interactions to link the same SNPs to target genes and observed a completely different set of unrelated biological processes for these genes (Figure 4.5C, right panel). To further characterize the biological relevance of target genes, we mined mouse knock-out data from the Mouse Genome

Informatics (MGI) database[173], which revealed that a statistically significant number of target genes resulted in a cardiovascular phenotype when knocked-out in the mouse (78 genes (22.4%), $P = 1x10^{-5}$, Figure 4.5D). Finally, we examined expression quantitative trait loci (eQTL) data from human left ventricle (LV) tissue and found that of the 1,999 LD SNPs in interactions, 410 (20.5%) corresponded to LV eQTLs; in comparison, only 12.2% of the full set of LD SNPs corresponded to LV eQTLs ($P < 0.00001$, Figure 4.5E). We next assessed whether eQTLs loop to their associated gene. For this analysis, we considered the full set of LV eQTLs, as the 410 LD SNP eQTLs represent too small of a proportion of the full set (< 0.1% of all LV eQTLs) to fully ascertain significance. On a genome-wide level, LV eQTLs in promoter-distal interactions were significantly more likely to loop to their associated gene than expected by chance ($P < 0.00001$, Figure 4.5F, left panel). Importantly, this significance decreased when LV eQTLs were analyzed with iPSC promoter interactions ($P = 0.035$, Figure 4.5F, right panel). Taken together, these results indicate that CM promoter interactions identify a subset of disease-relevant SNPs most likely to be functional and support the use of the CM map to assign distal CVD-associated SNPs to putative target genes.

Figure 4.5. CM promoter interactions link CVD GWAS SNPs to target genes. (A) Distribution of genomic distances separating SNP-target gene interactions (red line, median = 185 kb; black line, mean = 197 kb). **(B)** Pie chart showing the number of TSS's skipped for each SNP-target gene interaction (left) and the number of genes contacted by each SNP (right). **(C)** GO enrichment analysis for genes looping to LD SNPs using the CM promoter interaction data (left panel) or the iPSC promoter interaction data (right panel). **(D)** Proportion of target genes that result in a cardiovascular phenotype when knocked-out in the mouse (MGI database [173]), compared to a random control set. P-value calculated with a Z-test. **(E)** Proportion of GWAS LD SNPs that are eQTLs in left ventricle (LV) when considering either the full set of LD SNPs, or the subset that overlap CM promoter interactions. P-value calculated with Fisher's exact test. **(F)** Proportion of LV eQTLs (genome-wide) that map within a promoter interaction for the eQTL-associated gene (indicated by the red line). Random permutations were obtained by re-assigning each promoter's set of interactions to a new promoter and calculating the proportion of eQTLs in random interactions that interact with their eQTL-associated gene. Proportions only consider eQTLs that overlap a promoter-distal interaction. P-values calculated with a Z-test.



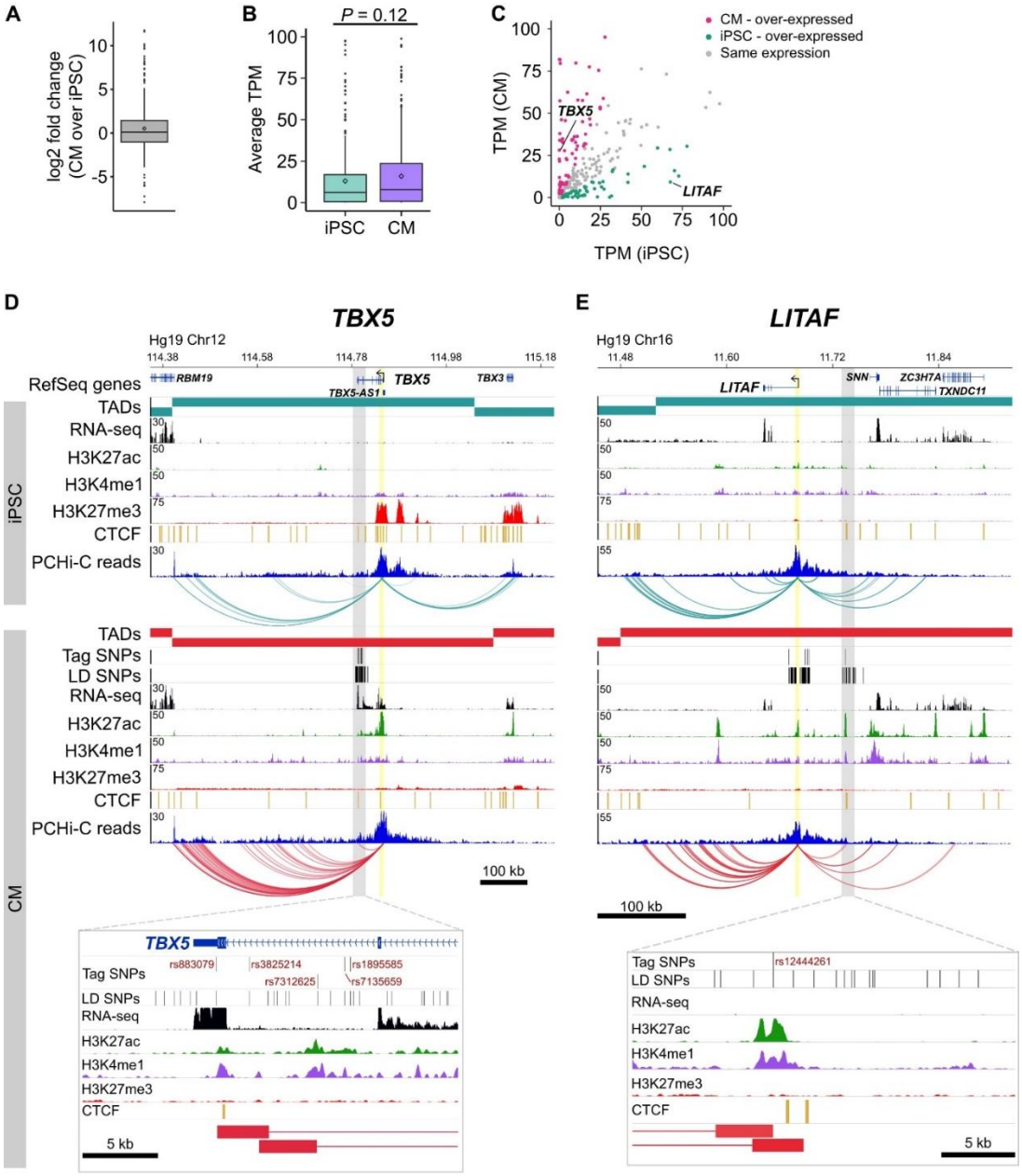
4.3.7 Using gene expression as a metric for interpreting disease-relevance of newly identified target genes

Based on an enrichment of target genes with known cardiac function, we next assessed whether expression level is an informative metric to further prioritize functional follow-up studies. We examined the expression level of the 347 target genes and found that they were moderately over-expressed in CMs compared to iPSCs (median log2 fold change = 1.08, mean log2 fold change = 1.44, mean TPM values were 40.6 in iPSCs and 60.1 in CMs, $P = 0.12$, Figures 4.6A and 4.6B). Although not significant, this result reflects the enrichment of known cardiac-related genes that interact with CVD loci. However, because a subset of target genes was over-expressed in iPSCs relative to CMs (Figure 4.6C), we predicted that gene expression level alone may be an insufficient metric to gauge the relevance of target genes to CVD biology. Indeed, we found that 21 of the 78 target genes (27%) that cause cardiovascular phenotypes when knocked-out in mice were overexpressed in iPSCs compared to CMs (Supplementary Link S4.8). This result indicates that putatively causal genes may not appear as obvious candidates based solely on gene expression data.

To illustrate this point, we highlight two genes: *TBX5*, a gene directly linked to cardiac arrhythmia (Figure 4.6D)[174,175], and *LITAF*, a gene that, until recently, had no obvious role in cardiac biology[176] (Figure 4.6E). Both genes formed long-range interactions to LD SNPs identified in arrhythmia GWAS, making both genes candidate functional targets of the GWAS associations. *TBX5*, which is over-expressed in CMs (Figure 4.6C), is the most likely target gene of the LD SNPs nearby based on the interaction data but also because of its known role in directing proper development of the cardiac conduction system. *LITAF*, on the other hand, was over-expressed in

iPSCs compared to CMs (Figure 4.6C), and was not known to contribute to cardiac function until a recent study identified this gene as a regulator of cardiac excitation in zebrafish hearts[176].

Figure 4.6 Characterizing target genes based on expression level. (A) Log2 fold change of the expression level of target genes in CMs compared to iPSCs (horizontal bar indicates median, 1.08; diamond indicates mean, 1.44). **(B)** Average TPM values of target genes in iPSCs and CMs (P = 0.12, Wilcoxon rank-sum test). Diamonds indicate the mean value (40.6 for iPSC, 60.1 for CM). **(C)** Comparison of average TPM values for target genes in CMs and iPSCs. See Supplementary Link S4.8 for full list of genes and TPM values. **(D,E)** Examples of genes looping to cardiac arrhythmia GWAS SNPs in CMs. **(D)** The *TBX5* gene interacts with a functionally validated arrhythmia locus[174]. **(E)** The *LITAF* gene interacts with a locus identified in[177]. Yellow highlighted region indicates the promoter; gray box and zoom panel show the promoter-interacting regions overlapping arrhythmia SNPs. For clarity, only interactions for the indicated promoter are shown.



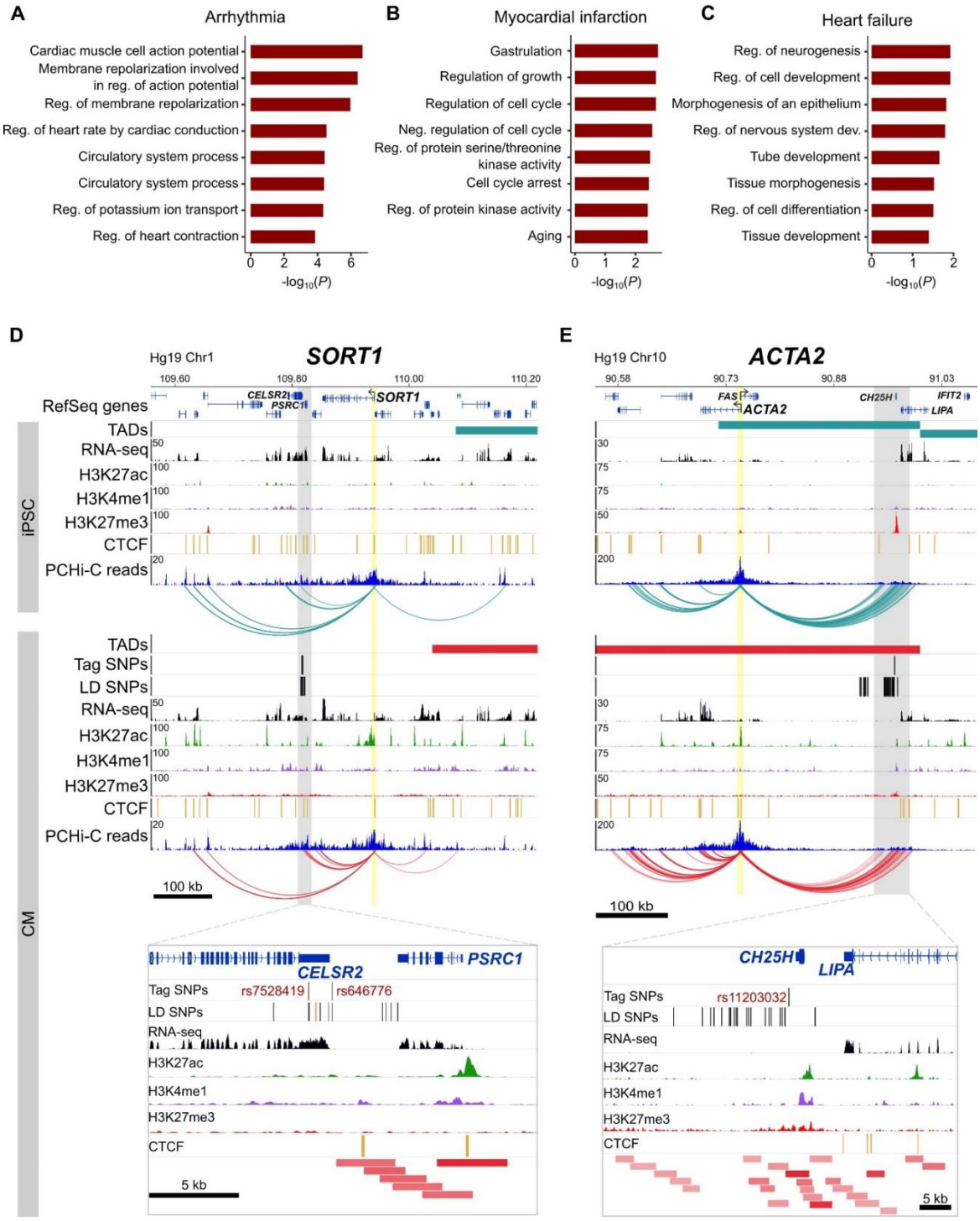
4.3.8 CM promoter interactions are informative to cardiovascular associations that do not directly involve cardiomyocytes

Because the three disease classes that we analyzed represent diverse pathologies, we predicted that the target genes identified for each class individually may relate to different biological processes. Specifically, we considered that cardiac arrhythmias—which directly result from defects in cardiomyocytes specialized for electrical conduction—may uncover the most cardiac-relevant target genes compared to heart failure and myocardial infarction, two CVDs that also involve non-cardiac systems. When broken down into the respective disease classes, we confirmed that the majority of the GO enrichment for cardiac terms was driven by the cardiac arrhythmia SNPs (Figure 4.7A), with terms directly related to the cardiac conduction system. Myocardial infarction (Figure 4.7B) and heart failure (Figure 4.7C) analyses uncovered a set of genes that were slightly enriched for regulation of growth and morphogenesis, respectively.

Despite these seemingly non-specific processes, each set of target genes contained important disease-relevant candidates. For example, one of the strongest associations for myocardial infarction lies in-between the *CELSR2* and *PSRC1* genes on chromosome 1p13, but a careful screen of genes whose expression was affected by the risk allele implicated the more distal *SORT1* gene[82]. *SORT1* encodes a sorting receptor that is expressed in many tissues and has been shown to act in the liver to regulate cholesterol levels[82,178]. Despite functioning in the liver, we identified multiple promoter interactions between *SORT1* and the myocardial infarction GWAS locus in CMs (Figure 4.7D), directly implicating *SORT1* as the target gene and lending further support to experimental validation of this locus as a *SORT1* enhancer[82]. Additionally, the *ACTA2* gene is located 220 kb away from the heart failure GWAS locus proximal to the *CH25H* and *LIPA* genes on chromosome 10q21[179] (Figure 4.7E). *ACTA2* encodes the smooth muscle cell-specific

actin protein and mutations in this gene have been shown to cause coronary artery disease, among other vascular diseases[180]. Despite its location at a considerable distance from the GWAS association, chromatin interactions provide an important level of evidence that *ACTA2* is a putative causal gene in the development of heart failure. Therefore, the CM interaction map is not only useful to interrogate diseases directly related to cardiomyocytes, as in the case of cardiac arrhythmias, but also aids interpretation of target genes that may act in non-cardiac tissues.

Figure 4.7 Relevance of CM promoter interactions for cardiac arrhythmia, myocardial infarction and heart failure. (A-C) Gene Ontology analysis for target genes looping to **(A)** cardiac arrhythmia SNPs, **(B)** myocardial infarction SNPs, and **(C)** heart failure SNPs. **(D)** The *SORT1* promoter loops to a distal myocardial infarction locus[82]. The rs12740374 SNP shown to disrupt a C/EBP binding site in[82] is colored red. **(E)** The *ACTA2* promoter loops to the 10q21 heart failure locus[179]. Zoom plots depict the full interacting region overlapping GWAS LD SNPs. For clarity, only interactions for the indicated gene are shown.



4.4 Discussion

Incomplete understanding of long-range gene regulation is a major roadblock in the translation of GWAS-associated loci to disease biology. Major challenges in this process include identifying putatively causal variants mapping within regulatory elements and functionally connecting these regulatory elements to their target genes. To delineate gene-regulatory interactions between CVD-associated SNPs and putative causal genes, we generated high-resolution maps of promoter interactions in human iPSCs and iPSC-derived CMs. We demonstrated that promoters interact with a diverse set of distal DNA elements in both cell types, including known enhancer sequences, which reflect cell identity and correspond to tissue-specific gene expression. To demonstrate the utility of the CM map, we linked 1,999 CVD-associated SNPs to putative causal target genes which identified both validated and potentially novel genes important for cardiovascular disease biology. To validate the biological relevance of our maps, we addressed several important features of long-range chromatin interactions in comparative analyses.

Promoters contact distal regions enriched for tissue-specific transcription factor motifs

Gene regulation by distant regulatory elements involves the bridging of linearly separated DNA sequences, for example between a promoter and its distal enhancers, through chromatin looping mechanisms[2]. In support of this model, we report an enrichment of tissue-defining transcription factor motifs in the distally interacting sequences of differentially expressed promoters both for CMs and iPSCs, providing an important level of evidence to validate the functional relevance of iPSC and CM interactions. One explanation for this enrichment is that our interaction maps are high resolution. We generated Hi-C libraries with the 4-bp cutter MboI, which generates fragments with an average size of 422 bp; this increased specificity of the captured

region likely leads to better resolution of the underlying enhancer sequence and, consequently, increased power to detect short transcription factor binding motifs.

Influence of active and repressive promoter interactions on gene expression level

The majority of capture Hi-C studies to date have reported that gene expression level correlates with enrichment for various histone marks. We observed the same trend in our data, with highly expressed genes exhibiting strong enrichment for looping to distal H3K4me1 and H3K27ac-marked regions, and lowly expressed genes exhibiting strong enrichment for looping to H3K27me3-marked regions. These data are consistent with a model in which the number of long-range interactions to enhancers or repressors additively contributes to gene expression level[49,146]. The forces that drive increased association between promoters and distal *cis*-regulatory elements are not completely understood and have been topics of investigation in the genome organization and chromatin biology fields for several years[181,182]. One possibility is that this increasing enrichment is driven by genomic compartmentalization of active and inactive chromatin. We showed that a gene's expression level correlates with the number of histone ChIP-seq peaks within a large window (300 kb) surrounding each promoter. Thus, highly expressed genes are more likely to contact active chromatin regions compared to lowly expressed genes, corresponding to the observed increasing enrichment of contacts and expression we and others have reported. This local increase in active or repressive chromatin may be one driving force underlying the expression level-dependent increase in association between promoters and *cis*-regulatory elements, akin to a phase separation-mediated model of enhancer-promoter interactions[183].

A promoter interaction map for cardiovascular disease genetics

We demonstrated several ways in which promoter interaction data can be used to better understand disease genetics, specifically addressing the major requirement for a high-resolution map of the gene-regulatory network in human cardiomyocytes. Although iPSC-derived CMs are known to be relatively immature and do not fully reflect the diverse structural and functional aspects of adult cardiac cells[184,185], the difficulty in obtaining pure sub-populations of primary cardiomyocytes with high integrity necessitates the use of an *in vitro* system. We showed that the CMs used in this study were highly pure and recapitulate known gene regulatory properties of primary cardiomyocytes. Because of this purity, we were able to integrate CVD-associated SNPs with CM promoter interactions with high confidence, assigning nearly 20% of the variants in high LD with these associations to 347 target genes.

Supporting the physiological relevance of CMs to the cardiac conduction system, we found that target genes were most relevant for GWAS loci associated with cardiac arrhythmias, in line with previous findings in immune cells that many target gene interactions were unique to relevant immune cell subtypes[78,146]. Our data also revealed that even for diseases whose etiology involves cell types other than cardiomyocytes, such as myocardial infarction and heart failure, we identified interactions involving loci associated with these diseases that recapitulate the enhancer-promoter interactions in non-cardiac cell types. As an example, we showed that a validated myocardial infarction locus interacts with the distal *SORT1* promoter in CMs even though this locus has been extensively characterized in the context of cholesterol metabolism in hepatocytes. Therefore, the promoter interactions we observe linking the disease locus to *SORT1* may represent tissue-invariant genome architecture, likely reflecting that genome organization in general is relatively stable[40,53,144]. While we advocate the use of the CM map for investigating gene regulatory

mechanisms of diseases related to cardiomyocyte biology, we also emphasize that, where identified, any interaction between a promoter and a putative disease-associated genomic region serves as an important level of evidence to prioritize that gene for future follow-up studies.

Limitations of the PCHi-C maps

The PCHi-C technique holds great promise to identify with high resolution and throughput all gene regulatory elements in any tissue or developmental stage of interest. However, due to technical and biological limitations, there are important caveats to PCHi-C that should be considered when interpreting the iPSC or CM interaction data. The most important caveat is that there are likely to be many false negatives, or "missing" interactions. Although the capture step greatly enriches for promoter-containing ligation fragments in a Hi-C library, the total landscape of promoter contacts in a population of cells is still under-sampled, even with a sequencing depth of ~400M reads per replicate conducted for this study. This is due to several factors, including the hybridization efficiency of each bait, ability to design sufficient baits per promoter, and the transient nature of many regulatory interactions. This latter issue is confounded by the distance-dependent effect on ligation frequency: as the distance between two fragments increases, the read-depth required to robustly identify that interaction also increases. The feasibility of deeper sequencing and modifications to computational pipelines will continue to improve the coverage and resolution of Hi-C data.

Additionally, because the CHiCAGO program does not incorporate TAD boundaries into its background model, it may slightly underestimate the expected number of reads corresponding to intra-TAD interactions which could lead to potential false positives. However, we note that there is a strong correspondence between TADs called on pre-capture Hi-C data and PCHi-C

interactions identified with CHiCAGO (Supplemental Figure S4.3); this suggests that accounting for TAD boundaries may only marginally improve our ability to identify significant interactions.

A final consideration is the interpretation of interactions involving inactive genes. Although most regulatory elements are thought of as activating, it is possible that long-range interactions may also contribute to gene silencing; this is supported by the observation that silent genes are enriched for long-range interactions to H3K27me3 marked regions (Figure 4.3D). Alternatively, silent genes may contact regulatory elements that are not active in the analyzed cell type or developmental stage; these may represent "pre-formed" loops between genes and their regulatory elements as characterized in[53].

Despite these limitations, the data sets we provide here represent a highly enriched set of ~350,000 and ~400,000 promoter interactions in iPSC and CMs, respectively; although there are likely missing interactions, the interactions that we did identify should be considered as very high confidence, as they were independently identified in at least two biological replicates and show strong signal of enrichment for known features of genome architecture and gene regulation. In conclusion, the promoter interaction maps we generated in this study represent important resources for any investigation into the gene regulatory mechanisms underlying cardiovascular disease traits. The list of candidate regulatory variants and their target genes may serve as an entry point for several hypotheses related to CVD GWAS, and can be readily tested in experimental settings. To provide both the iPSC and CM maps as an accessible resource, we have hosted the full set of data presented in this study at the WashU EpiGenome Browser[186], accessible at the following link: http://epigenomegateway.wustl.edu/browser/?genome=hg19&publichub=Lindsey. Additionally, we provide the significant PCHi-C interaction files used in all analyses in the Supplemental File

(Supplemental Files S4.1 and S4.2); these can be applied to future multi-omics analyses of gene regulation and disease genetics.

4.5 Methods

4.5.1 Tissue culture of iPSCs

We used the Yoruban iPSC line 19101, kindly provided by the laboratory of Yoav Gilad. This iPSC line was reprogrammed from lymphoblastoid cells as part of a previous study, where it was shown to differentiate into all three germ layers, displayed a normal karyotype, and expressed markers characteristic of pluripotency[187]. iPSCs were grown in Essential 8 (E8) Medium (Thermo Fisher #A1517001) supplemented with 1X Penicillin-Streptomycin (Pen/Strep, Gibco) on Matrigel-coated tissue culture dishes (Corning #354277). Cells were passaged when they were ~80% confluent using enzyme-free dissociation solution (30mM NaCl, 0.5mM EDTA, 1X PBS minus Magnesium and Calcium) and maintained in E8 Medium with 10μM Y-27632 dihydrochloride (Abcam #ab120129) for 24 hours. Medium was replaced daily. iPSC cultures routinely tested negative for mycoplasma contamination using the Universal Mycoplasma Detection Kit (ATCC #30-1012K).

4.5.2 Cardiomyocyte differentiation

Cardiomyocyte differentiations were based on the protocol of Burridge *et al.*[150] with modifications described in Banovich *et al.*[187]. iPSCs were expanded in 60 mm dishes in E8 media until they reached 60-70% confluency at which time the differentiation was started (day 0). On day 0, E8 media was replaced with 10mL of basic heart media/12μM GSK-3 inhibitor CHIR-99021 trihydrochloride (Tocris #4953)/Matrigel overlay [basic heart media: RPMI 1640 minus L-glutamine (HyClone #SH30096.01) with 1X GlutaMax (Life Technologies #11879020)

supplemented with 1X B27 minus insulin (Thermo Fisher #A1895601) and 1X Pen/Strep; Matrigel overlay was accomplished by dissolving Matrigel in 50mL basic heart media at a concentration of 0.5X according to the lot-specific dilution factor]. After 24 hours (day 1), the GSK-3 inhibitor was removed by replacing media with 10 mL basic heart media. On day 3, media was replaced with 10 mL basic heart media supplemented with 2 μM Wnt-C59 (Tocris #5148). On day 5 (48 hours later), media was replaced with 10 mL basic heart media. On day 7, cells were washed once with 1X PBS and then 15 mL basic heart media was added. Media was replaced every other day in this way until day 15 at which time cardiomyocytes were selected for by replacing basic heart media with 10mL lactate media (RPMI 1640 minus D-glucose, plus L-glutamine (Life Technologies #11879020), supplemented with 0.5 mg/mL recombinant human albumin (Sigma 70024-90-7), 5 mM sodium DL-lactate (Sigma 72-17-3), 213 μg/mL L-ascorbic acid 2-phosphate (Sigma 70024-90-7) and 1X Pen/Strep). Lactate media was replaced every other day until day 20 at which point cardiomyocytes were harvested. Cells from successful differentiations exhibited spontaneous beating around days 7-10.

Cardiomyocytes were harvested by washing once with 1X PBS followed by incubation in 4 mL TrypLE (Life Technologies 12604-021) at 37°C for 5 minutes. After incubation, 4 mL lactate media was added to the TrypLE and a 1 mL pipet was used to dislodge cells. Cells were strained once with a 100μM strainer and then once with a 40μM strainer. Cells were pelleted at 500xg and then resuspended in PBS and counted. For each batch of differentiation, 5 million cells were taken for promoter-capture Hi-C and 1 million cells were taken for RNA-seq. To assess purity, 2 million cells were taken for flow cytometry analysis using an antibody for cardiac Troponin T (BD Biosciences 564767). All cells used in downstream experiments were at least 86% Troponin T positive (Supplemental Figure S4.1A). We carried out three independent differentiations of the

same iPSC line and generated promoter-capture Hi-C and RNA-seq libraries in iPSCs and CMs from each triplicate.

4.5.3 Promoter capture Hi-C

Crosslinking cells

iPSCs or cardiomyocytes were harvested from tissue culture dishes and counted. Cells were resuspended in 1X PBS at a concentration of 1 million cells/mL and 37% formaldehyde was added to a final concentration of 1%. Crosslinking was carried out for 10 minutes at room temperature on a rocking platform. Glycine was added to a final concentration of 0.2 M to quench the reaction. The cells were pelleted, snap frozen in liquid nitrogen and stored at -80°C until ready for Hi-C processing.

in situ Hi-C

We prepared all promoter capture Hi-C libraries in one batch using three crosslinked pellets of 5 million cells for both iPSCs and iPSC-derived cardiomyocytes, representing three independent cardiomyocyte differentiations. The *in situ* Hi-C step was performed as in Rao *et al.*[143] with a single modification in which NEBNext reagents from the NEBNext Multiplex Oligos for Illumina kit were used (NEB #E7335S) instead of Illumina adapters, following the manufacturer's instructions. Hi-C libraries were amplified directly off of T1 beads (Life Technologies #65602) using NEBNext primers and 6 cycles of PCR.

Promoter capture – probe design and generation

Hi-C capture probes were designed to target four MboI restriction fragment ends (120 bp) near the TSS of protein coding RefSeq genes[188] mapped to hg19 in the UCSC Genome Browser[189].

To select restriction fragments, we only kept MboI restriction fragments longer than 200 bp and overlapping 10 kb around a RefSeq TSS. TSSs closer than 1 kb from each other were excluded, as their interactions were likely to be captured by the other RefSeq TSS. The four MboI restriction fragments ends closest to each RefSeq TSS were selected as putative probes. The 120 bp sequences were submitted to Agilent's SureDesign proprietary software for probe selection, which can slightly shift the location and remove probes. In total, we ordered a library of 77,476 single-stranded DNA oligos from CustomArray, Inc. (www.customarrayinc.com). Each oligo consisted of the sequence 5'-ATCGCACCAGCGTGTN$_{120}$CACTGCGGCTCCTCA-3'[190] where N$_{120}$ represents the 120 nucleotides adjacent to the MboI cut site. The complete list of oligo probes and their corresponding gene name is provided in Supplemental File S4.9.

The oligos arrived as a pool containing 1000 ng of material. We used 16 ng of the oligo pool in a PCR reaction to make them double stranded using primers 5′-CTGGGAATCGCACCAGCGTGT-3′ (Primer A), and 5′-CGTGGATGAGGAGCCGCAGTG-3′ (Primer B) as in[190]. The PCR reaction was cleaned using AMPure XP beads (Agencourt #A6388) and eluted with 20µl of water. To add the full T7 promoter to the 5' end of the oligos, a second PCR reaction was carried out using 10ng of the cleaned-up first-round PCR product with the forward primer 5'-GGATTCTAATACGACTCACTATAGGGATCGCACCAGCGTGT-3' (Primer A T7). We purified the PCR product corresponding to 176 bp using a Qiagen gel extraction kit (#28704). To generate biotinylated RNA baits, we performed *in vitro* transcription on the double-stranded library using the MEGAshortscript T7 Transcription Kit (Thermo Fisher #AM135) with Biotin-16-dUTP (Sigma #11388908910). After DNase treatment the transcription reaction was cleaned using the MEGAclear kit (Thermo Fisher #AM1908) and eluted with 50µl elution buffer. We confirmed the correct bait size on a denaturing gel.

Promoter capture – hybridization with Hi-C library

To isolate promoter-containing fragments from the whole-genome *in situ* Hi-C library, we hybridized the biotinylated RNA bait pool with the Hi-C library as follows. A mix containing 500ng of the Hi-C library, 2.5µg of human Cot-1 DNA (Invitrogen #15279-011), 2.5µg of salmon sperm DNA (Invitrogen #15632-011), 0.5µl blocking primer P5 (IDT #1016184), and 0.5µl blocking primer P7 (IDT #1016186) was heated for 5 min. at 95°C, held at 65°C and mixed with 13µl pre-warmed hybridization buffer (10X SSPE, 10X Denhardt's, 10 mM EDTA and 0.2% SDS) and a 6 µl pre-warmed mix of 500ng of the biotinylated RNA bait and 20U SUPERase-In (Thermo Fisher #AM2694). The hybridization mix was incubated for 24h at 65°C. To isolate captured fragments, we prepared 500ng of streptavidin-coated magnetic beads (Dynabeads MyOne Streptavidin T1, Thermo Fisher #65601) in 200µl of Binding buffer (1M NaCl, 10mM Tris-HCl pH 7.5, 1mM EDTA). The hybridization mix was added to the Streptavidin beads and rotated for 30 minutes at room temperature. The beads containing the captured Hi-C fragments were washed with 1X SSC, 0.1% SDS for 15 minutes at room temperature, followed by three washes (10 min each) at 65°C with 0.1X SSC/0.1% SDS. After the final wash, the beads were resuspended in 22µl of water and proceeded to post-capture PCR. The PCR reaction was performed with 11µl of the "capture Hi-C beads" and 8 cycles of amplification. An AMPure XP bead purification was used to clean the PCR reaction and DNA was quantified using the QuantiFluor dsDNA System (Promega #E2670) and a High Sensitivity Bioanalyzer. Final capture Hi-C libraries were subjected to 100bp paired-end sequencing on an Illumina HiSeq 4000 machine. Read count summaries are provided in Supplementary Link S4.9.

4.5.4 Interaction calling

We used HiCUP v0.5.9[191] to align and filter Hi-C reads (total and filtered read counts are presented in Supplementary Link S4.9). Unique reads were given to CHiCAGO version 1.2.0[151] and significant interactions were called with default parameters. In this study, we focused exclusively on *cis*-interactions as the evidence that *trans*-chromosomal interactions contribute to gene expression regulation is limited. CHiCAGO reports interactions for each captured restriction fragment; to summarize interactions by gene, we considered the interval spanning all captured fragments (i.e. the set of probes spanning each TSS) as the promoter region ("merged TSS"). This means the promoter regions created have variable lengths. In cases where multiple genes were annotated to the same promoter region, we report the interaction for each gene individually. This annotation allowed us to perform gene-level analyses, for example based on expression level. We removed this redundancy as necessary, for example in motif enrichment analyses of the promoter-interacting fragments. Using the "merged TSS" interaction files, we filtered interactions to retain those that mapped within 1 kb of each other in at least two replicates. Specifically, we extended each promoter-interacting fragment by 1 kb on each end and then used BEDTools[97] pairToPair functionality to identify interactions where both ends matched across replicates. To identify cell type-specific interactions, we required that the interaction (with the 1 kb extension) was not present in any of the three replicates of the other cell type. The number of read-pairs per promoter and the corresponding number of significant interactions identified is presented in Supplementary Link S4.9. The TAD analyses, motif enrichment, ChIP-seq peak enrichment, and eQTL analyses (related to Figures 4.1, 4.2, 4.3 and 4.5) were conducted with fragment-level interactions (no 1kb extension). The GWAS SNP analyses were conducted with 1kb-extended interactions, as we aimed to be as inclusive as possible when linking CVD SNPs to target genes.

PCHi-C interactions, TADs, RNA-seq, publicly available ChIP-seq, and GWAS SNPs are hosted by the WashU EpiGenome Browser[186] as a public track hub. This can be accessed by going to http://epigenomegateway.wustl.edu/browser/. The public hub ("A promoter interaction map for cardiovascular disease genetics") can be found under the Human Hg19 browser.

4.5.5 4C-style plots

To generate the by-gene read counts displayed in the genome-browser figures, all read-pairs mapping to captured MboI fragments for a given promoter were summed across replicates. Specifically, we summed reads for each MboI fragment where the read was part of a paired-read that mapped to a bait for the given gene. The arcs that are displayed underneath the 4C-style plot represent significant interactions that were identified in at least two replicates as detailed above in "Interaction calling".

4.5.6 TAD analysis

To identify TADs, we pooled reads across replicates for each cell type using the pre-capture Hi-C data (600M reads for iPSC and 733M reads for CM) and used HiCUP v0.5.9[191] to align and filter Hi-C reads. HOMER v4.8.3[90] was used to generate normalized interaction matrices at a resolution of 40 kb and then TopDom v0.0.2[192] was used with a window size w=10 to identify topological domains, boundaries and gaps. We only considered domains for the analyses in this paper. We considered a promoter capture Hi-C interaction to be "intra-TAD" if the entire span of the interaction was fully contained in a single domain. "Inter-TAD" interactions are defined as interactions where each end maps to a different domain.

4.5.7 A/B compartments

The program runHiCpca.pl from the HOMER[90] v4.8.3 package was used to call A/B compartments with -res 50000 for both whole-genome and capture Hi-C data.

4.5.8 RNA-seq

Total RNA was extracted from flash-frozen pellets of 1 million cells using TRI Reagent (Sigma #T9424) and a homogenizer followed by RNA isolation and clean-up using the Direct-zol RNA Kit (Zymo Research #11-331). RNA-seq libraries were generated with the Illumina TruSeq V2 kit (Illumina, RS-122-2001) and 1μg of RNA, following manufacturer's instructions. Libraries were made from RNA isolated from three independent iPSC-CM differentiations (triplicates of iPSC and of cardiomyocytes). Libraries were sequenced on an Illumina HiSeq 4000.

Gene counts were quantified with Salmon 0.7.2[120] and imported with tximport 1.2.0[121] into DESeq2 1.12.4[122] to call differentially expressed genes. A minimum 1.5-fold-difference between CMs and iPSC triplicates and a minimum adjusted P-value of 0.05 were required to select differentially expressed genes for downstream analyses. TPMs (transcripts per million) were also estimated by Salmon. Because the samples clearly clustered according to their known tissues of origin (Supplemental Figure S4.2A), no correction for batch effects was performed.

4.5.9 H3K27ac ChIP-seq for comparison with Epigenome Roadmap samples

We performed ChIP-seq on 2.5 million cells each for iPSCs and CMs using H3K27ac antibodies (Wako #306-34849). Briefly, cells were crosslinked with 1% formaldehyde for 10 minutes at room temperature, quenched with 0.2M glycine for 5 minutes, pelleted and snap-frozen in liquid nitrogen. Cells were lysed in Lysis Buffer 1 (50 mM HEPES-KOH, pH 7.5, 140 mM NaCl, 1 mM EDTA, 10% glycerol, 0.5% NP-40, 0.25% Triton X-100). Crosslinked chromatin was

sheared to an average size of 300 bp using a Bioruptor with 30" on/30" off at high setting and then incubated overnight at 4ºC with 1 µg antibody. Dynabeads M-280 Sheep Anti-Mouse IgG (ThermoFisher #11201D) were used to pull down chromatin and ChIP DNA was eluted and prepared for sequencing using the NEBNext Ultra II DNA Library prep kit (NEB #E7645S). ChIP-seq reads were aligned with Bowtie 2-2.2.3[96] and peaks were called with HOMER[90] v4.8.3 on unique reads with mapping quality > 10 using the –region and –style histone parameters. Significant peaks were overlapped with H3K27ac peaks from Epigenome Roadmap samples which demonstrated high concordance between matched tissue types (Supplemental Figure S4.2C,D). Because we performed a low level of sequencing, we did not identify as many peaks as the Roadmap samples. Therefore we used Roadmap ChIP-seq data in all of our analyses.

4.5.10 Gene Ontology analysis

The human Gene Ontology (GO) associations of GO terms[193] to genes and the GO database were downloaded on January 22, 2016 from http://geneontology.org/gene-associations. GO terms were associated with RefSeq genes via gene symbols. Using the GO annotation graph, all parent terms were assigned to the terms annotated to a gene. A hypergeometric test was used to calculate the statistical significance of the difference of the number of genes associated with a given GO term in a particular gene set and the universe of all RefSeq genes ($P < 0.05$). P-values were corrected with the R package p.adjust function using the "fdr" method.

For two of the GWAS disease groups (heart failure and myocardial infarction), the list of genes looping to LD SNPs included many histone genes. This is because there is a tag SNP located in the middle of a histone gene cluster (containing > 30 histone genes located close together) in each case. After expanding the tag SNP to all SNPs in LD, many of the histone genes in that cluster looped to the LD SNPs, resulting in a high representation of these genes in the final gene list. The

resulting Gene Ontology enrichment analysis gave terms relating to nucleosome and chromatin organization because of this over-representation. We therefore chose to remove these genes from the final gene lists of heart failure and myocardial infarction target genes.

4.5.11 Motif analysis

The program findMotifsGenome.pl from the HOMER[90] v4.8.3 package was used with –size given parameter to identify overrepresented motifs in the distal (non-promoter) interacting sequences of promoter interactions. As stated above, this analysis was performed on fragment-level promoter-interacting sequences.

4.5.12 Histone ChIP-seq enrichment analysis

We obtained publicly available ChIP-seq data in the form of processed peak calls for H3K27ac, H3K4me1 and H3K27me3 from the Roadmap Epigenomics Project[13], and for CTCF from ENCODE[194] (Supplemental File S4.10). We only considered peaks that mapped outside of the captured region of promoters to ensure our results were not driven by the strong peak signal over most promoters. As a proxy for iPSCs, we used data from the H1 embryonic stem cell line and for CMs we used data from Left Ventricle tissue. We grouped genes into five expression categories based on the average TPM values: group 1 (0 TPM), group 2 (TPM 0-3), group 3 (TPM 3-25), group 4 (TPM 25-150) and group 5 (TPM >150) and for each group of genes, we calculated the enrichment for promoter interactions to overlap a given feature. To calculate enrichment of interactions overlapping an epigenetic feature, we compared the observed proportion of MboI fragments in significant interactions overlapping a feature to the proportion of random MboI fragments overlapping the feature. Specifically, we randomly selected MboI fragments from a set that excluded fragments mapping within captured regions (promoters) or within unmappable

genomic regions (gaps). The number of randomly selected fragments matched the number of interacting fragments considered for the analysis. We performed 100 iterations of overlapping random fragments with a feature and report the average fold-enrichment. We refer to this method of enrichment as a "genome-wide" background model because for each gene expression group, the observed proportion of fragments containing a peak is compared to randomly selected fragments from the whole genome.

To calculate the correlation between expression and histone ChIP-seq peak density, we calculated the Spearman's rank correlation between the expression value for each gene (the average TPM value) and the number of peaks mapping within 300 kb of each gene TSS. We only considered genes with at least one significant interaction in the respective cell type to allow for generalizations to the enrichment analysis presented in Figure 4.3.

4.5.13 GWAS analysis

We compiled genome-wide significant SNPs associated with GWAS for cardiac arrhythmia, heart failure, and myocardial infarction from the NHGRI-EBI database (http://www.ebi.ac.uk/gwas/); see Supplemental File S4.7 for list of terms used to identify specific GWAS. We expanded each set of SNPs to all SNPs in high LD ($r^2 > 0.9$) using phase 3 data of the 1,000 genomes project[195] (Supplemental File S4.3). For each lead SNP from the GWAS we analyzed, we selected a 100 kb interval centered on the SNP (SNP +/- 50kb). For each 100 kb interval, Tabix[196] was used to retrieve genotypes. We then used PLINK[197] v1.90p on phase 3 data from the 1,000 genomes project[195] (ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502, v5a) to select SNPs in LD ($r^2 > 0.9$) with the tag SNP and a minimum allele frequency of 0.01. We only included the populations primarily studied in the GWASs: CEU (central European), ASW (African American) and JPT (Japanese). We assigned all SNPs in promoter-distal interactions to their

interacting gene(s) ("target genes") using cardiomyocyte promoter capture Hi-C data. We did not require the SNP to map to regions associated with open chromatin or enhancer marks as these types of data are highly cell-type specific and we did not wish to exclude SNPs in regions that may be active in non-assayed cell types.

We note that one major GWAS for dilated cardiomyopathy was not included in the NHGRI-EBI database[198], likely because there is an error obtaining the online methods of the paper. After careful inspection of the study, we concluded that the GWAS met the NHGRI-EBI criteria and included the associations from that study in our analysis. A complete list of all studies used in this analysis can be found in Supplemental File S4.8.

4.5.14 MGI analysis

To calculate enrichment of target genes to cause cardiovascular phenotypes when deleted in mice (Mouse Genome Informatics database), we randomly selected 347 genes from the list of starting genes (i.e. genes with at least one promoter-distal interaction in CMs, meaning it could be a target gene), and calculated the proportion that caused a cardiovascular phenotype in mice. We performed this randomized selection for 100 iterations to generate the randomized (expected) values. Random genes were not required to be expressed, as the set of target genes contains genes that are not expressed. P-value was calculated with a Z test.

4.5.15 eQTL analysis

For eQTLs used in comparisons with GWAS variants and Hi-C interactions, we used the set of GTEx v7 eQTLs identified as significant in the left ventricle of the heart[199]. eQTLs were called significant if $q<0.05$ after false discovery rate correction[200]. We only considered promoter-

distal eQTLs that were at least 10 kb from their associated gene to allow for that eQTL to map to an interaction with its associated gene.

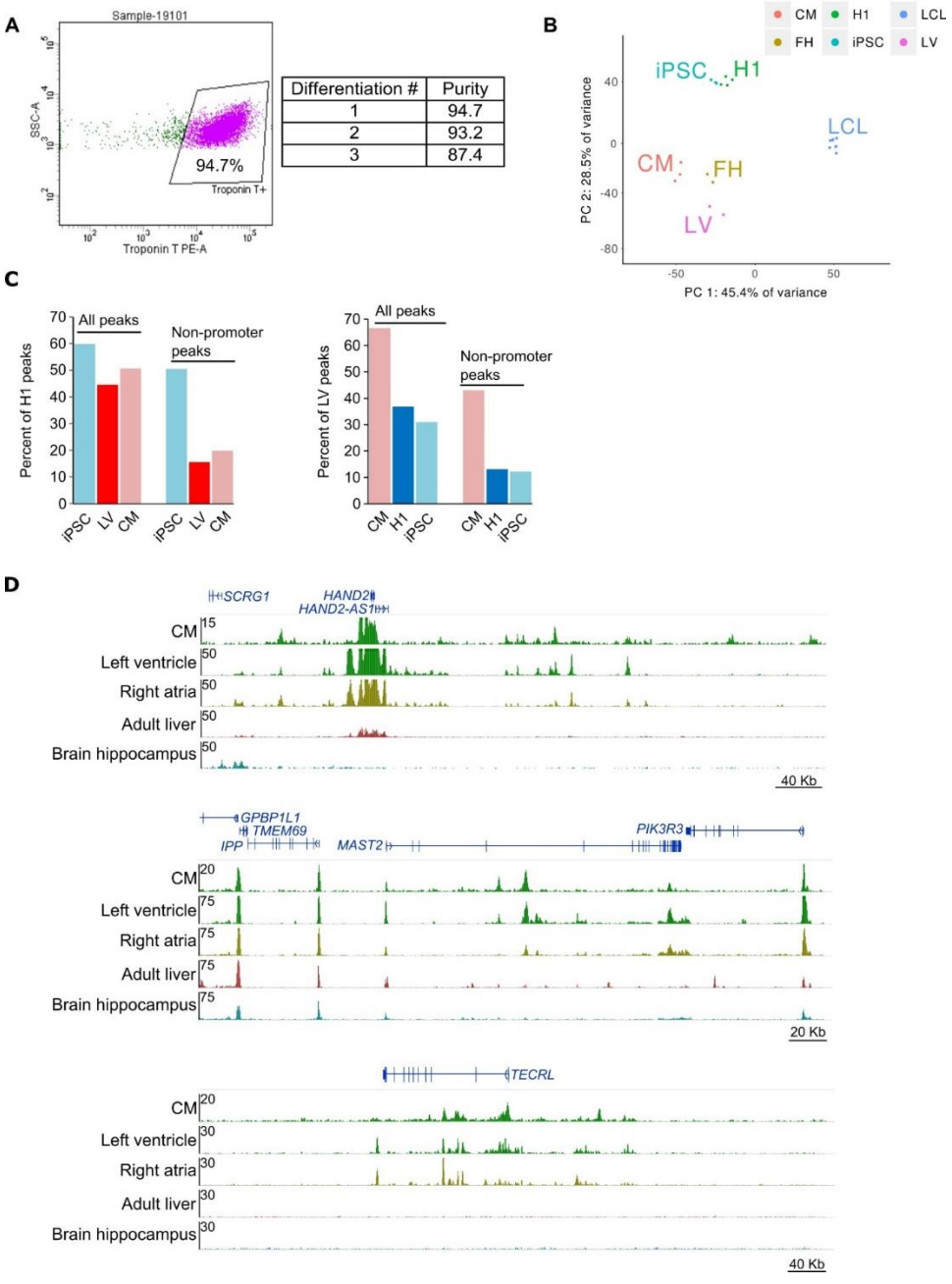
To calculate enrichment for eQTLs to loop to their associated gene, we used a background model whereby each promoter's set of interactions were re-mapped to a different promoter, keeping the distance and strand orientation consistent. We performed this re-mapping of all promoter interactions 1000 times and calculated the proportion of all eQTLs that mapped to interactions for their eQTL-associated gene in each permutation. We either used the CM interactions or the iPSC interactions with the same set of left ventricle eQTLs to compare cell-type specificity of the promoter interaction data.

4.5.16 Data Availability

Raw and processed sequencing data are provided at ArrayExpress through accession numbers E-MTAB-6014 (Hi-C) and E-MTAB-6013 (RNA-seq).

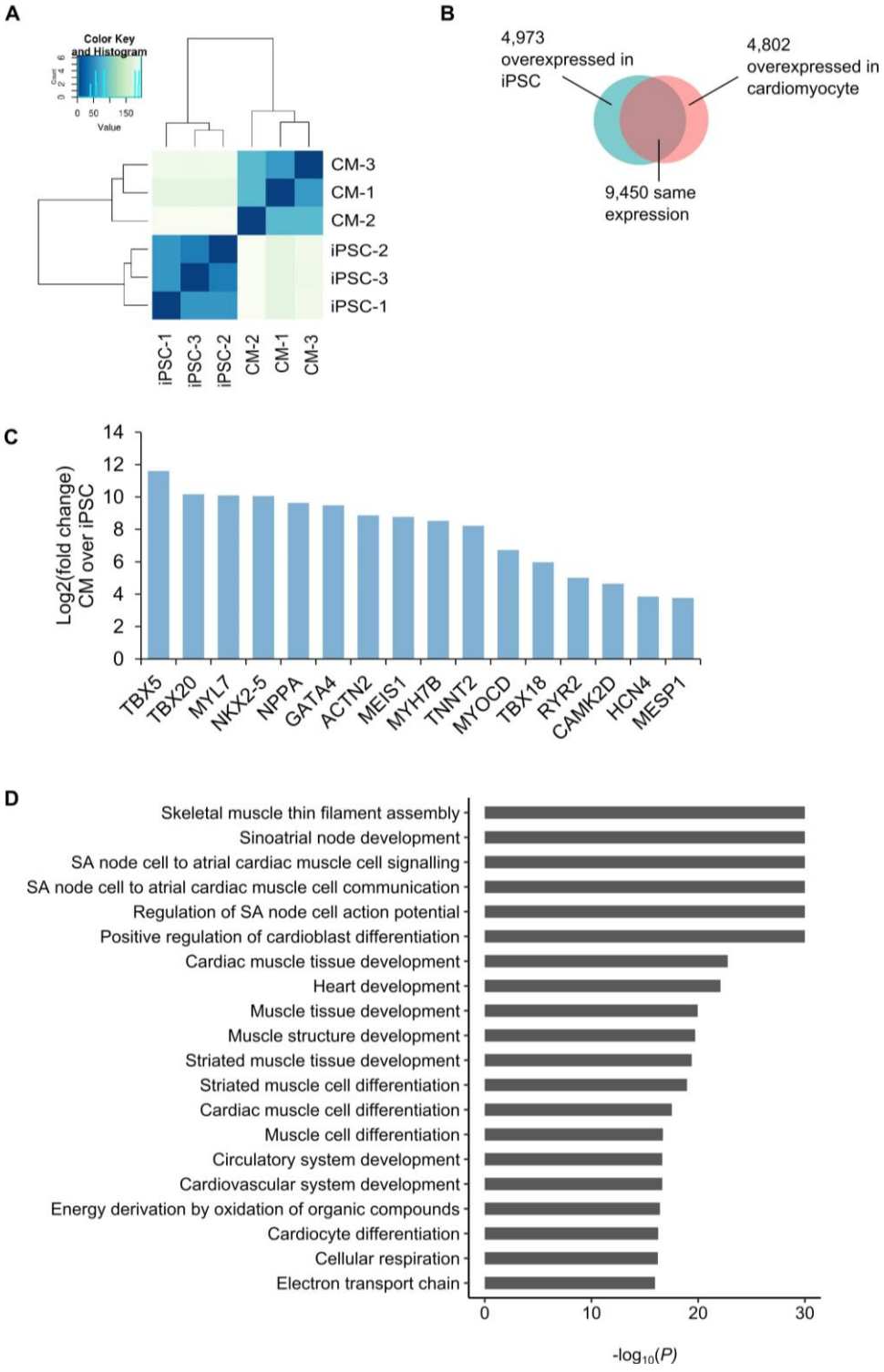
4.6 Appendix E: Supplementary Figures

Supplemental Figure S4.1 Quality control of iPSC-CMs (A) Flow cytometry of iPSC-derived cardiomyocytes. Representative image of flow data for cardiomyocytes (left) and percent cardiac troponin T (cTnT) positive for each differentiation (right). Cells were first gated on live/dead and then on cTnT staining.

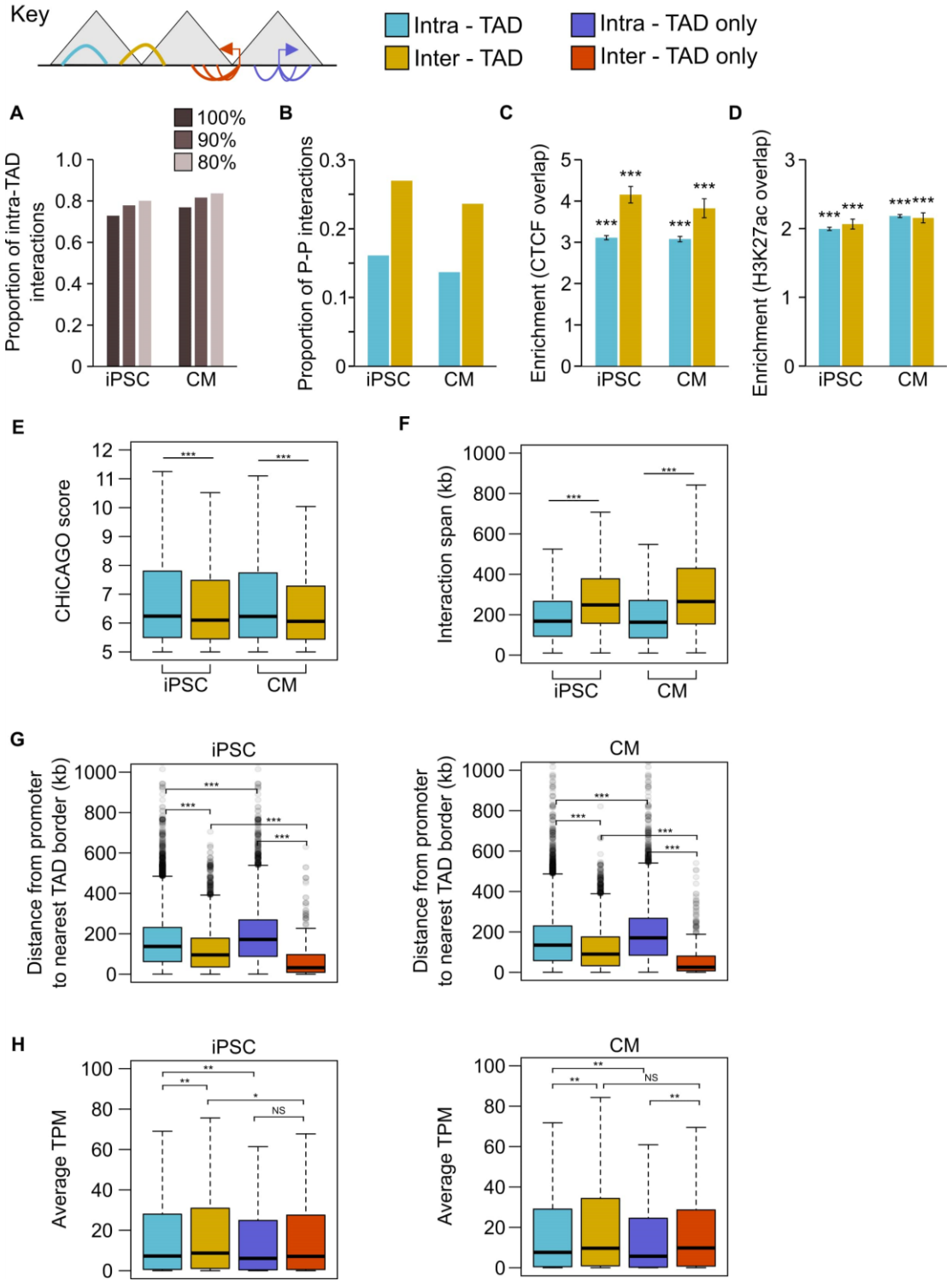


Supplemental Figure S4.1, continued. (B) Principle component analysis of RNA-seq data in iPSCs and CMs along with H1 embryonic stem cells, left ventricular cells (LV), fetal heart cells (FH), and lymphoblastoid cell line cells (LCL). LCLs cluster independently from iPSC and CM, indicating that iPSCs were faithfully reprogrammed. **(C)** Percentage of Epigenome Roadmap H3K27ac ChIP-seq peaks overlapping iPSC and CM H3K27ac peaks. Overlaps for all peaks and only non-promoter peaks are shown. LV, left ventricle; H1, H1 embryonic stem cell line. **(D)** Three genome browser snap-shots displaying the epigenetic landscape in CMs compared to left ventricle, right atria, adult liver and brain hippocampus from the Epigenome Roadmap.

Supplemental Figure S4.2 Analysis of RNA-seq in iPSCs and iPSC-CMs (A) Cluster analysis of RNA-seq data from each triplicate of iPSC and CM. **(B)** Number of genes differentially expressed in each cell type. **(C)** Selected genes overexpressed in CMs relative to iPSCs. **(D)** Gene Ontology enrichment analysis of the biological processes associated with the 4,802 genes overexpressed in cardiomyocytes.

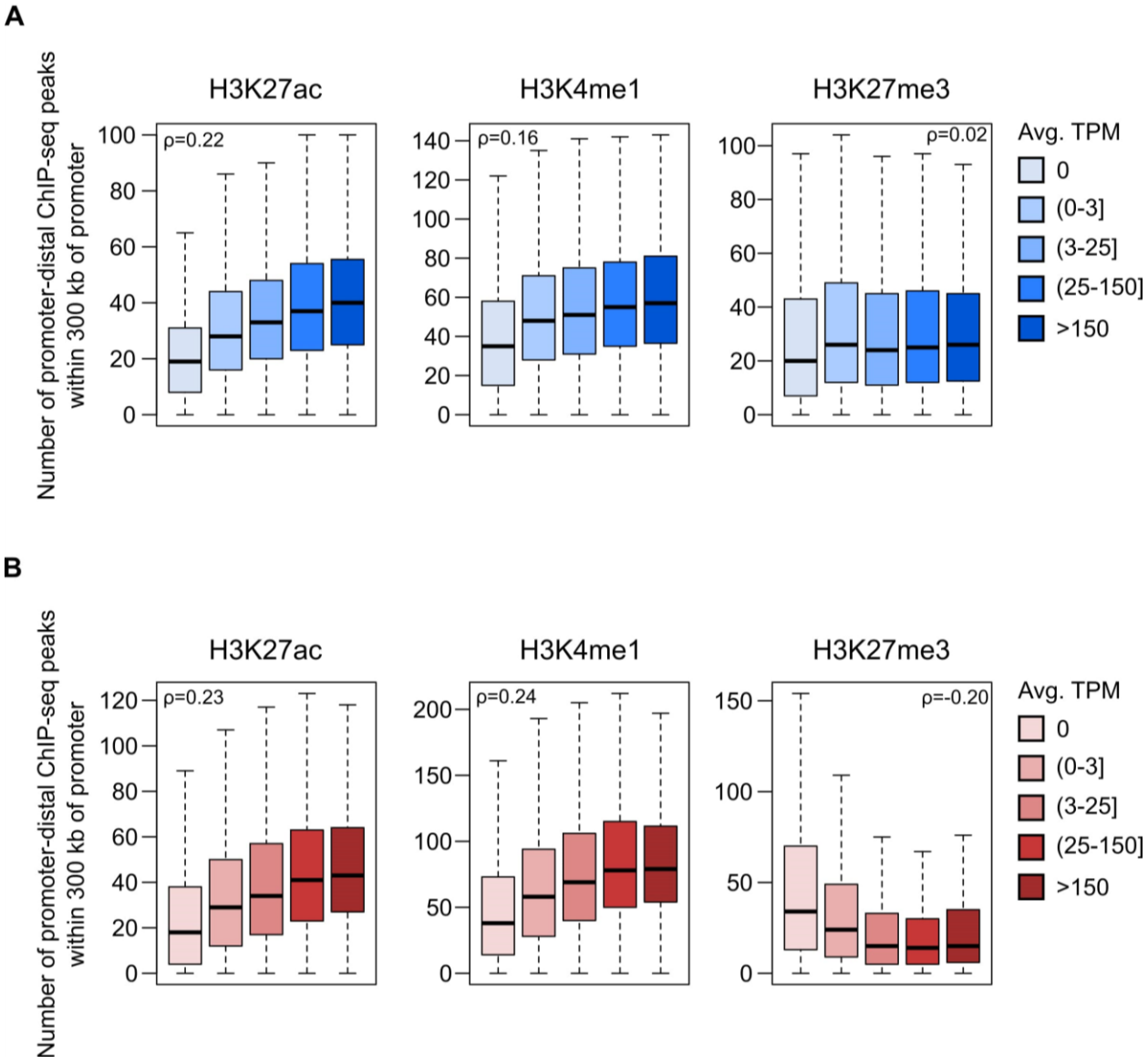


Supplemental Figure S4.3 Analysis of PCHi-C interactions in the context of TADs In this analysis, interactions were classified as intra-TAD (both ends of the interaction fully within a single TAD) or inter-TAD (each end of the interaction is in a different TAD). Interactions falling partially or wholly within TAD "boundaries" or "gaps" as defined by TopDom were omitted (see section 4.5 Methods).

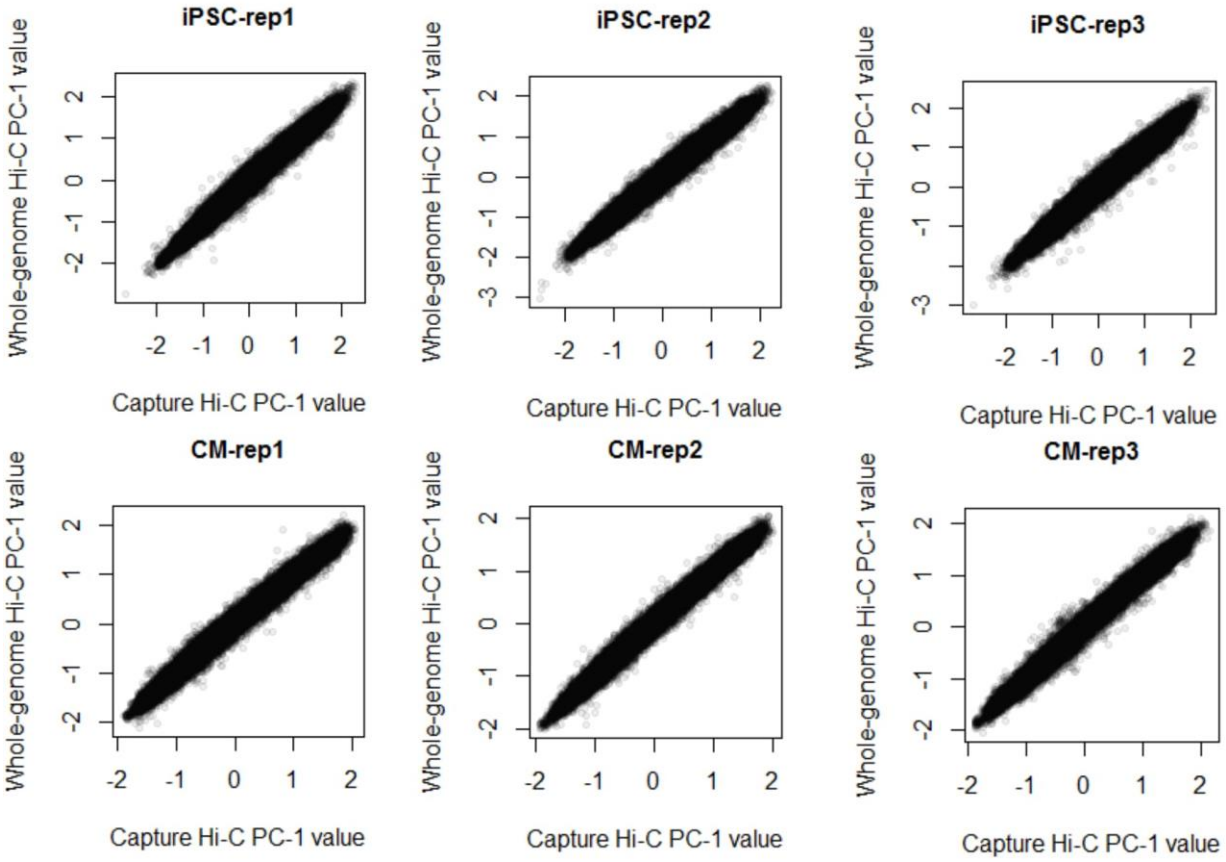


Supplemental Figure S4.3, continued. (A) Proportion of interactions that are intra-TAD at different cut-offs. All analyses used interactions that were 100% within a TAD. **(B)** Proportion of promoter-promoter interactions in the set of intra-TAD and inter-TAD interactions. **(C,D)** Fold enrichment for intra-TAD and inter-TAD interactions to overlap CTCF **(C)** or H3K27ac peaks **(D)**. Only promoter-distal ChIP-seq peaks were analyzed. ***$P < 2.2x10^{-16}$, Z-test. **(E)** CHiCAGO score and **(F)** interaction span of intra- vs. inter-TAD interactions. ***$P < 2.2x10^{-16}$, Wilcoxon rank-sum test. **(G,H)** Considering promoters with an intra-TAD interaction, an inter-TAD interaction, or exclusively intra-TAD or inter-TAD interactions: **(G)** distance from the promoter TSS to the nearest TAD boundary and **(H)** average TPM value of the promoter. ***$P < 2.2x10^{-16}$, **$P < 0.01$, *$P < 0.05$, NS = not significant, Wilcoxon rank-sum test.

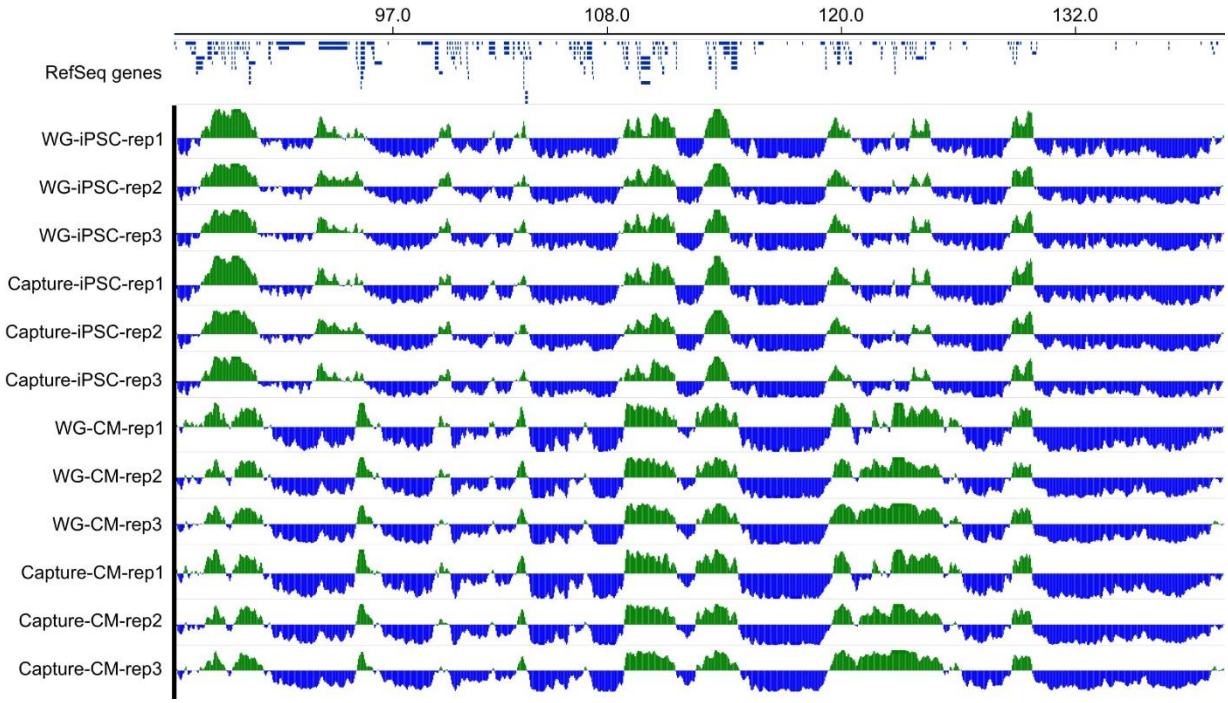
Supplemental Figure S4.4 Correlation between the number of histone ChIP-seq peaks within 300 kb of promoters and gene expression level. Number of promoter-distal histone ChIP-seq peaks within 300 kb of promoters in iPSC **(A)** and CM **(B)**. Spearman's rho (ρ) was calculated on the full set of promoter expression values/peak counts for all promoters with at least one significant interaction in the respective cell type (12,926 genes for iPSC and 13,555 genes for CM; see section 4.5 Methods). Data are grouped by expression category to emphasize the trend. Horizontal bars indicate the median for each expression category. All correlation estimates are significant at $P < 2.2x10^{-16}$ except for H3K27me3 in iPSCs ($P = 0.06$).



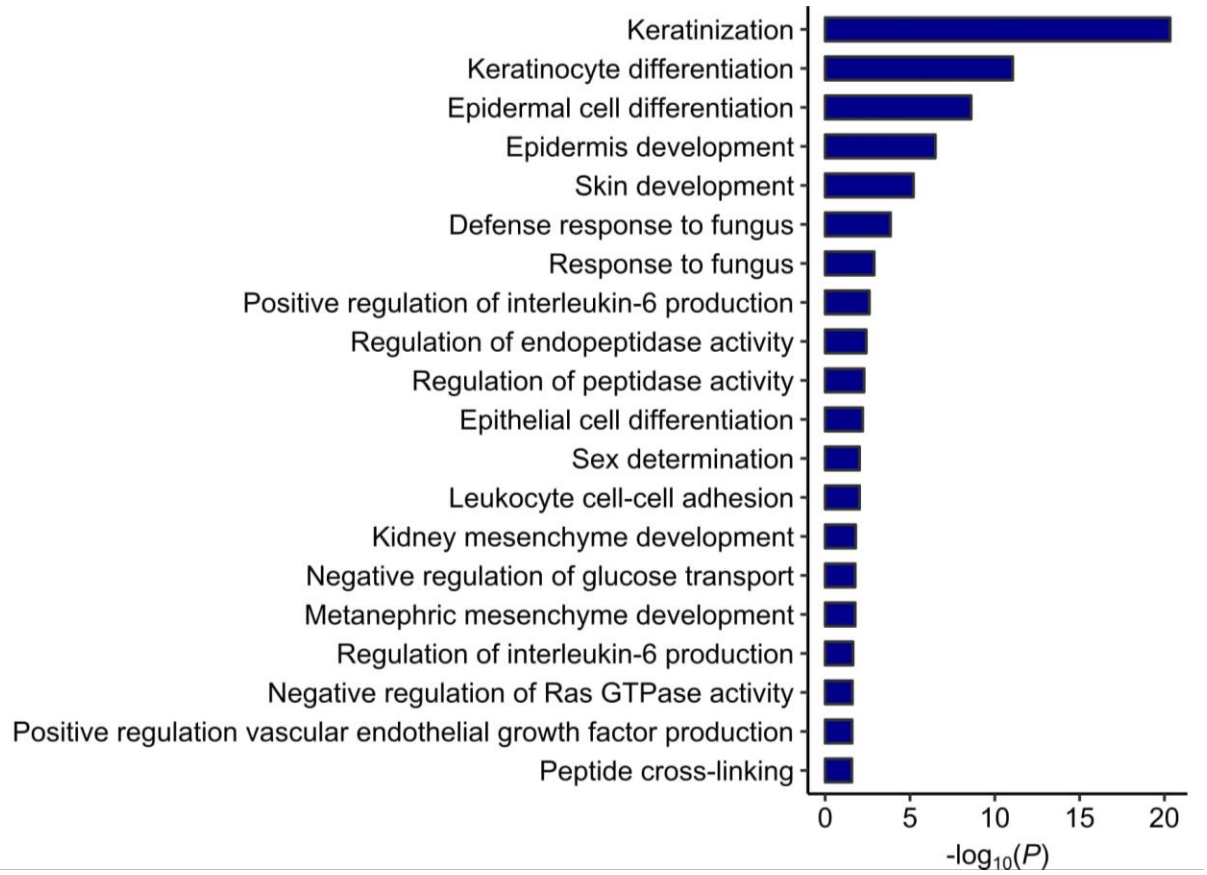
Supplemental Figure S4.5 Comparison of A/B compartments in Hi-C and PCHi-C Correlation between the A/B compartment score (principle component analysis of interaction data, PC-1) in whole-genome Hi-C (y-axis) and promoter capture Hi-C (x-axis) in iPSCs (top) and CMs (bottom). Spearman's $\rho > 0.98$, $P < 2.2 \times 10^{-16}$ in all cases.



Supplemental Figure S4.6 Example of A/B compartments Genome browser snapshot of a ~53 Mb region on chromosome 4 showing A/B compartments in all three replicates of iPSCs and CMs using both whole-genome (WG) and promoter capture Hi-C data.



Supplemental Figure S4.7 GO analysis on the genes switching from active A compartments in iPSCs to inactive B compartments in CMs. GO analysis on the genes switching from active A compartments in iPSCs to inactive B compartments in CMs.



4.7 Appendix F: Web links for supplemental files

Supplemental File S4.1 PCHi-C interactions for iPSC
https://doi.org/10.7554/eLife.35788.019
Supplemental File S4.2 PCHI-C interactions for CM https://doi.org/10.7554/eLife.35788.020
Supplemental File S4.3 CVD SNPs https://doi.org/10.7554/eLife.35788.021
Supplemental File S4.4 HOMER motif analysis for the distal interacting regions of promoter interactions https://doi.org/10.7554/eLife.35788.022
Supplemental File S4.5 Gene Ontology enrichment output
https://doi.org/10.7554/eLife.35788.023
Supplemental File S4.6 Gene Ontology input gene lists
https://doi.org/10.7554/eLife.35788.024
Supplemental File S4.7 GWAS terms used to compile studies
https://doi.org/10.7554/eLife.35788.025
Supplemental File S4.8 GWAS summary table https://doi.org/10.7554/eLife.35788.026
Supplemental File S4.9 Hi-C read information https://doi.org/10.7554/eLife.35788.027
Supplemental File S4.10 Public datasets used https://doi.org/10.7554/eLife.35788.028

CHAPTER 5: SUMMARY AND CONCLUSIONS

The overarching goal of my dissertation research was to improve our understanding of how the genome encodes and carries out its function in regulating gene expression, with the hope that this work would increase our ability to interpret the genetic basis of human disease. To this end, I used multiple orthogonal approaches to investigate the genome from a gene-regulatory perspective, with three specific aims: improve an assay designed to identify regulatory elements, study the function of two of the most conserved enhancer elements in the vertebrate genome, and generate a promoter interaction map to provide a framework for identifying and prioritizing target genes implicated in CVDs.

Improving a method to survey gene regulatory elements

In Chapter 2, I developed an improvement to the widely-used ATAC-seq assay that tackles the problem of mitochondrial read contamination which has been reported in ATAC-seq experiments[93]. This work used CRISPR/Cas9 to selectively cleave DNA fragments originating from the mitochondrial genome and prevent their contribution to the final sequencing library. This treatment dramatically reduced the high abundance of mitochondrial sequencing reads which resulted in an increased sensitivity to detect regulatory elements such as enhancers and promoters, reduced the overall cost of sequencing, and led to increased reproducibility of peaks as a result of the higher quality of libraries. We have made the key reagent, the mitochondrial gRNA library, freely available to the research community in the hopes that this approach will improve the ATAC-seq assay in those cell types that tend to be susceptible to mitochondrial read contamination, particularly immune cell types.

Using enhancer deletions to analyze the function of ultraconserved elements

In Chapter 3, I studied the role of two ultraconserved elements using mouse models. UCEs have fascinated researchers since their original identification in 2004, largely because of the extreme level of non-coding nucleotide conservation characteristic of many UCEs[22]. Support that UCEs act as critical regulatory elements stems from studies showing that a large proportion of UCEs drive reporter gene activity in specific tissues during embryonic development[24,100]. However, the fact that germline deletion of seven UCEs failed to show any developmental or physiological phenotype indicates that these elements may not function as stereotypical enhancers, as was originally thought[26–28].

The two UCEs that I focused on, UCE3 and UCE5, are unique in that they share a core ~300 bp sequence with each other as well two other UCEs located in a paralogous region of a different chromosome. This level of nucleotide sequence conservation strongly suggests that these elements play a critical role in the regulation and/or function of the *Irx* genes which share the syntenic block containing each UCE pair[103]. In order to explicitly test the necessity of these two elements in *Irx3* and *Irx5* gene regulation, I used CRISPR/Cas9 genome editing to delete each element from the mouse genome and predicted that one or both deletions would affect *Irx3*/*Irx5* gene expression in the hypothalamus and consequently cause a body weight phenotype[74]. Surprisingly, although the UCE5 deletion caused a body weight phenotype, this was not accompanied by the predicted change in gene expression, indicating that we likely missed the time-frame within which the UCE deletion affected *Irx3*/*Irx5* gene expression, or alternatively that the body weight phenotype reflects a general disruption to organismal health that is not necessarily caused by mis-regulation of *Irx3*/*Irx5* expression. Deletion of the UCE3 element did not cause a molecular or physiological phenotype.

This work was designed to test the gene-regulatory function of two specific UCEs which were thought to be outliers with respect to the other UCEs due to their unique level of intra-element conservation. In light of my results, it appears that the *Irx* UCEs are not obviously different from other UCEs, at least to the extent that UCEs have been analyzed in deletion studies. For example, only 3 out of 8 UCEs deleted to date (excluding the present study) resulted in a physiological phenotype (hs121 and hs122 from Dickel *et al.*[28]), and only 2 out of 8 caused a gene expression change (M280 from Nolte *et al.*[27] and hs122 from Dickel *et al.*[28]). When considered together with our study, these results suggest that the true function underlying UCE conservation may stem from a mechanism that is not yet fully elucidated, but that likely extends beyond stereotypical enhancer function. Such a mechanism would explain the tissue-specific transcriptional activity observed for a large number of UCEs but would be consistent with an absence of gene expression changes or a physiological phenotype when an element is fully removed from the genome.

One potential mechanism of function is that UCEs influence gene expression indirectly by acting as transcription factor binding site hubs which may aid in regulating the local concentration and subsequent activity of transcriptional proteins to ensure proper expression of the developmentally important genes nearby. For example, Viturawong *et al.* demonstrated that UCEs contain dense, overlapping binding sites for dozens of protein factors, mostly tissue-specific transcription factors[102], and that the majority of UCE nucleotides contributed to multiple overlapping binding sites, which increases the selection constraint on each of those positions. These results indicate that UCE sequences may serve as docking sites for dozens of protein factors which is consistent with their strong transcriptional activation activity and their propensity to remain nucleosome-depleted. One possibility is that this dense binding platform helps recruit and retain important transcriptional regulators during sensitive periods of embryonic development. The

Irx3 and *Irx5* genes encode critical transcription factors and their expression should be tightly regulated; the nearby UCE3 and UCE5 elements may act as "protein sinks" to increase and/or regulate the local concentration of transcriptional regulators that act on *Irx3* and *Irx5*. Furthermore, there has been a recent surge in evidence showing that many nuclear factors can form biomolecular condensates in liquid-like droplets which may contribute to the regulation of their transcriptional activity[201–203]. Thus, UCEs in general may serve as nucleating sites along chromosomes to aid in the formation of certain regulatory droplets near genes whose expression is highly sensitive to perturbations in expression levels. It may be expected that deletion of one of these elements should dramatically alter the expression of the nearby target gene, however the fact that there are often multiple UCEs near these genes indicates that redundancy exists to protect against this possibility. Indeed, Dickel *et al.* observed that deletion of two UCEs was often required to obtain a developmental or gene expression phenotype[28]. Moreover, the existence of the UCEs themselves may serve a purely redundant role to ensure transcriptional robustness in the absence of a primary regulatory input in the form of a stereotypical enhancer, which are replete in the *Irx* gene clusters[103]. As the work I presented only deleted one UCE at a time, it remains to be determined whether deletion of both UCE3 and UCE5 would have led to a more dramatic gene-regulatory alteration.

A final consideration is the use of whole-scale deletions to study the role of a putative regulatory element. We assume that a UCE, or any enhancer for that matter, functions primarily through biding sequence-specific transcription factors. As such, removing the entire repertoire of binding sites (i.e. the entire element) is hypothesized to alter target gene activity. However, we know that many "back-up" mechanisms exist to ensure robustness of gene expression in the absence of a primary input[204]; therefore, an orthogonal approach to assay the contribution of a

regulatory element to gene expression would be to mutate the element. This approach would leave the element intact in the genome, thereby potentially neutralizing the activity of redundant enhancers, yet would alter the binding sites in a way that is perhaps more likely to impact target gene expression.

In summary, I showed through germline deletion studies that two of the most conserved elements in the vertebrate genome are largely dispensable for proper development, although it is worth noting that the significant body weight phenotype caused by the UCE5 deletion may provide the selective constraint required to maintain sequence conservation. Although several lines of evidence suggest that these elements function as enhancers for the *Irx3* and *Irx5* genes, we did not find evidence that deletion of either UCE disrupted *Irx3*/*5* expression in the hypothalamus, where we know from previous studies that *Irx3*/*Irx5* plays a critical role in regulating body weight homeostasis[74]. These results are consistent with what has been reported for other UCEs and suggests that the true mechanism by which UCEs function is yet to be fully elucidated. By considering alternative modes of action, such as acting as protein hubs instead of long-range enhancers, and by employing alternative approaches to disrupt UCEs, such as mutation instead of deletion, it may be possible to gain further insight into their elusive functions in vertebrate life.

A 3D framework to study the genetic basis of cardiovascular disease

In the final chapter of my thesis, I sought to develop a framework within which to interpret the hundreds of loci associated with cardiovascular diseases. Currently, there is a major gap in our ability to transform GWAS findings into interpretable, actionable metrics that may be used to design therapeutic interventions for complex disease. This is largely due to the time-consuming yet necessary functional experimental research that is required to unambiguously resolve the phenotypic consequences of a disease-associated locus. For these reasons, despite nearly 4,000

GWAS and tens of thousands of trait-associated loci, there are only a handful of "success stories" whereby the gene responsible for the association is known[205]. Identifying these target genes is the first step towards understanding the biology of the genetic basis of complex disease.

I addressed this problem in Chapter 4 of my thesis, where I generated high resolution promoter capture Hi-C data in human cardiomyocytes which can be used to systematically identify the most likely target gene(s) of genetic variants associated with cardiovascular disease. In total, I functionally connected 1,999 CVD-associated SNPs to 347 target genes in CMs. I validated these pairings in several ways. First, I showed that the target genes are enriched for genes with known roles in cardiovascular biology. It is important to note that target genes were identified purely based on their interactions with distal CVD-associated variants and not gene expression, demonstrating a connection between the promoter interaction landscape in CMs and the CVD genetic landscape. Second, I showed that target genes are enriched for causing cardiovascular phenotypes when deleted in the mouse. This analysis revealed that there are dozens of genes with potentially unknown roles in CVD biology, opening the door for numerous functional follow-up studies. Third, I highlighted the example of *SORT1* which has been previously identified as the functional target of one of the strongest myocardial infarction GWAS loci[82]; *SORT1* interacts with this locus in my data, validating the previous functional work and the interaction data that I generated. Finally, I showed that promoter interactions were in general enriched for connecting left ventricular eQTLs to their associated gene(s) in CMs, supporting the use of promoter interactions to correctly pair cis-regulatory elements to target genes.

It is now widely appreciated that the genetic basis of complex disease is rooted in gene regulation, not protein-coding alterations. Given that the field of long-range gene regulation is now flourishing, thanks in large part to the development of 3C-based technologies, it is not surprising

that investigators are seeking approaches to interpret GWAS results in the context of the three-dimensional genome. During the five years I have been working on my PhD, more than a dozen different studies were published which used genome organization data, including PCHi-C, to link disease loci to target genes, demonstrating that this approach is proving useful for interpreting putative functions of GWAS loci.

Despite the promises of these approaches and the progress that has been made in such a short time, there is still a long way to go before translating GWAS findings to actionable metrics becomes feasible. In my view, two key issues stand out: First, it is important to verify beyond correlative relationships that a disease-implicated variant is causal. This means unambiguously demonstrating that the variant and/or the larger (presumably cis-regulatory) region containing the variant actively controls a target gene's expression in a disease-relevant tissue/developmental stage. The emergence of new technologies will prove critical in achieving this goal. For example, massively parallel reporter assays (MPRAs) enable high throughput testing of allele-specific enhancer activity, and CRISPR-based enhancer perturbation methods such as CRSIPR-activation or CRISPR-interference will aid identification of gene(s) regulated by distal variants[206]. Combining these approaches with genome organization maps will result in the identification of genes whose expression level variation is a causal factor in the development of disease.

The second key issue is to understand *how* and *why* alterations in specific gene expression levels contribute to disease. This is a much less tractable problem because it should ideally require an ability to experimentally perturb gene expression levels on small scales in order to study the physiological effects of small changes in expression levels of many genes over time. It may never become feasible to experimentally investigate gene regulatory changes at this scale in living organisms, however theories such as the omnigenic model[65] posit that one need not necessarily

understand the contribution of each gene to a trait if we have an understanding of the gene *network* that is really at play. One value of this framework is that there are far fewer networks compared to genes, and networks are more likely to be therapeutically targeted. Nevertheless, I would argue that understanding the physiological consequences of small changes in gene expression is a worthwhile endeavor as indicated from recent reports which show that newly developed drugs are twice as likely to succeed if there is genetic data supporting the drug target[207,208].

In conclusion, the fields of human genetics and gene regulation are combining forces to tackle the problem of understanding complex disease. My work here contributed to these efforts by generating 3D promoter interaction maps in a cell type relevant for cardiovascular disease, adding to the growing repertoire of genome organization maps in diverse cell types. As more layers of gene regulatory information are generated, we will be able to begin to synthesize the myriad molecular inputs that collectively control gene expression, and perhaps it will become possible to perturb systems in a way to study the physiological changes caused by subtle alterations in gene expression. This trajectory holds promise for gaining deeper insights into the genetic and molecular basis of complex human disease.

BIBLIOGRAPHY

1. van Heyningen, V. & Bickmore, W. Regulation from a distance: long-range control of gene expression in development and disease. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **368,** 20120372 (2013).

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

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

4. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci.* **107,** 21931–21936 (2010).

5. Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459,** 108–12 (2009).

6. Zentner, G. E., Tesar, P. J. & Scacheri, P. C. Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. *Genome Res.* **21,** 1273–83 (2011).

7. Calo, E. & Wysocka, J. Modification of Enhancer Chromatin: What, How, and Why? *Mol. Cell* **49,** 825–837 (2013).

8. Boyle, A. P. *et al.* High-Resolution Mapping and Characterization of Open Chromatin across the Genome. *Cell* **132,** 311–322 (2008).

9. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10,** 1213–8 (2013).

10. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr. Protoc. Mol. Biol.* **109,** 21.29.1-9 (2015).

11. Montefiori, L. *et al.* Reducing mitochondrial reads in ATAC-seq using CRISPR/Cas9. *Sci. Rep.* **7,** 2451 (2017).

12. Stamatoyannopoulos, J. A. *et al.* An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biology* **13,** 418 (2012).

13. Roadmap Epigenomics Consortium, A. *et al.* Integrative analysis of 111 reference human

epigenomes. *Nature* **518,** 317–30 (2015).

14. Corces, M. R. *et al.* The chromatin accessibility landscape of primary human cancers. *Science* **362,** eaav1898 (2018).

15. Cusanovich, D. A. *et al.* A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell* **174,** 1309–1324.e18 (2018).

16. Hardison, R. C. Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet.* **16,** 369–72 (2000).

17. Ahituv, N., Rubin, E. M. & Nobrega, M. A. Exploiting human-fish genome comparisons for deciphering gene regulation. *Hum. Mol. Genet.* **13,** R261–R266 (2004).

18. Weirauch, M. T. & Hughes, T. R. Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same. *Trends Genet.* **26,** 66–74 (2010).

19. Hare, E. E., Peterson, B. K., Iyer, V. N., Meier, R. & Eisen, M. B. Sepsid even-skipped Enhancers Are Functionally Conserved in Drosophila Despite Lack of Sequence Conservation. *PLoS Genet.* **4,** e1000106 (2008).

20. Blow, M. J. *et al.* ChIP-seq identification of weakly conserved heart enhancers. *Nature Genetics* (2010). doi:10.1038/ng.650

21. Farley, E. K., Olson, K. M., Zhang, W., Rokhsar, D. S. & Levine, M. S. Syntax compensates for poor binding sites to encode tissue specificity of developmental enhancers. *Proc. Natl. Acad. Sci. U. S. A.* **113,** 6508–13 (2016).

22. Bejerano, G. *et al.* Ultraconserved Elements in the Human Genome. *Science (80-.).* **304,** 1321–1325 (2004).

23. Woolfe, A. *et al.* Highly Conserved Non-Coding Sequences Are Associated with Vertebrate Development. *PLoS Biol.* **3,** e7 (2004).

24. Pennacchio, L. a *et al.* In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444,** 499–502 (2006).

25. Kikuta, H. *et al.* Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Res.* **17,** 545–555 (2007).

26. Ahituv, N. *et al.* Deletion of ultraconserved elements yields viable mice. *PLoS Biol.* **5,** e234

(2007).

27. Nolte, M. J. *et al.* Functional analysis of limb transcriptional enhancers in the mouse. *Evol. Dev.* **16,** 207–223 (2014).

28. Dickel, D. E. *et al.* Ultraconserved Enhancers Are Required for Normal Development. *Cell* **172,** 491–499.e15 (2018).

29. Kosak, S. T. *et al.* Subnuclear compartmentalization of immunoglobulin loci during lymphocyte development. *Science (80-.).* **296,** 158–162 (2002).

30. Jhunjhunwala, S. *et al.* The 3D Structure of the Immunoglobulin Heavy-Chain Locus: Implications for Long-Range Genomic Interactions. *Cell* **133,** 265–279 (2008).

31. Williams, R. R. E. *et al.* Neural induction promotes large-scale chromatin reorganisation of the Mash1 locus. *J. Cell Sci.* **119,** 132–40 (2006).

32. Hewitt, S., High, F., Reiner, S., Fisher, A. & Merkenschlager, M. Nuclear repositioning marks the selective exclusion of lineage-inappropriate transcription factor loci during T helper cell differentiation. *Eur. J. Immunol.* **34,** 3604–3613 (2004).

33. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing Chromosome Conformation. *Science (80-.).* **295,** 1306–1311 (2002).

34. Zhao, Z. *et al.* Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat. Genet.* **38,** 1341–1347 (2006).

35. Dostie, J. *et al.* Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. *Genome Res.* **16,** 1299–1309 (2006).

36. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326,** 289–93 (2009).

37. Lanctôt, C., Cheutin, T., Cremer, M., Cavalli, G. & Cremer, T. Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. *Nat. Rev. Genet.* **8,** 104–115 (2007).

38. Miele, A. & Dekker, J. Long-range chromosomal interactions and gene regulation. *Mol. Biosyst.* **4,** 1046–57 (2008).

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

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

41. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485,** 376–380 (2012).

42. Nora, E. P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485,** 381–385 (2012).

43. Harmston, N. *et al.* Topologically associating domains are ancient features that coincide with Metazoan clusters of extreme noncoding conservation. *Nat. Commun.* **8,** 441 (2017).

44. Kumasaka, N., Knights, A. J. & Gaffney, D. J. High-resolution genetic mapping of putative causal interactions between regions of open chromatin. *Nat. Genet.* **51,** 128–137 (2019).

45. Lupiáñez, D. G. *et al.* Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. *Cell* **161,** 1012–1025 (2015).

46. Franke, M. *et al.* Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* (2016). doi:10.1038/nature19800

47. Symmons, O. *et al.* The Shh Topological Domain Facilitates the Action of Remote Enhancers by Reducing the Effects of Genomic Distances. *Dev. Cell* **39,** 529–543 (2016).

48. Tsujimura, T. *et al.* A Discrete Transition Zone Organizes the Topological and Regulatory Autonomy of the Adjacent Tfap2c and Bmp7 Genes. *PLoS Genet.* **11,** e1004897 (2015).

49. Schoenfelder, S. *et al.* The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Res.* **25,** 582–597 (2015).

50. Mifsud, B. *et al.* Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.* **47,** 598–606 (2015).

51. Freire-Pritchett, P. *et al.* Global reorganisation of *cis* -regulatory units upon lineage commitment of human embryonic stem cells. *Elife* **6,** (2017).

52. Javierre, B. M. *et al.* Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* **167,** 1369–1384 (2016).

53. Ghavi-Helm, Y. *et al.* Enhancer loops appear stable during development and are associated with paused polymerase. *Nature* **512,** 96 (2014).

54. Paliou, C. *et al.* Preformed Chromatin Topology Assists Transcriptional Robustness of Shh during Limb Development. *bioRxiv* 528877 (2019). doi:10.1101/528877

55. Platt, J. L. *et al.* Capture-C reveals preformed chromatin interactions between HIF-binding sites and distant promoters. *EMBO Rep.* **17,** 1410–1421 (2016).

56. Calhoun, V. C. & Levine, M. Long-range enhancer-promoter interactions in the Scr-Antp interval of the Drosophila Antennapedia complex. *Proc. Natl. Acad. Sci.* **100,** 9878–9883 (2003).

57. Nolis, I. K. *et al.* Transcription factors mediate long-range enhancer-promoter interactions. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 20222–7 (2009).

58. Bulger, M. & Groudine, M. Functional and Mechanistic Diversity of Distal Transcription Enhancers. *Cell* **144,** 327–339 (2011).

59. Dekker, J. & Mirny, L. The 3D Genome as Moderator of Chromosomal Communication. *Cell* **164,** 1110–1121 (2016).

60. Deng, W. *et al.* Reactivation of Developmentally Silenced Globin Genes by Forced Chromatin Looping. *Cell* **158,** 849–860 (2014).

61. Chen, H. *et al.* Dynamic interplay between enhancer–promoter topology and gene activity. *Nat. Genet.* 1 (2018). doi:10.1038/s41588-018-0175-z

62. Lettice, L. A. *et al.* A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.* **12,** 1725–1735 (2003).

63. Smemo, S. *et al.* Regulatory variation in a TBX5 enhancer leads to isolated congenital heart disease. *Hum. Mol. Genet.* **21,** 3255–3263 (2012).

64. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101,** 5–22 (2017).

65. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169,** 1177–1186 (2017).

66. 1000 Genomes Project Consortium, A. *et al.* A global reference for human genetic variation.

Nature **526,** 68–74 (2015).

67. Aguet, F. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550,** 204–213 (2017).

68. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. (2017). doi:10.1016/j.ajhg.2017.06.005

69. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. SUPP. *Science* **337,** 1190–5 (2012).

70. Nicolae, D. L. *et al.* Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. *PLoS Genet.* **6,** e1000888 (2010).

71. Frayling, T. M. *et al.* A Common Variant in the FTO Gene Is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity. *Science (80-.).* **316,** 889–894 (2007).

72. Church, C. *et al.* Overexpression of Fto leads to increased food intake and results in obesity. *Nat. Genet.* (2010). doi:10.1038/ng.713

73. Fischer, J. *et al.* Inactivation of the Fto gene protects from obesity. *Nature* (2009). doi:10.1038/nature07848

74. Smemo, S. *et al.* Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* **507,** 371–5 (2014).

75. Dryden, N. H. *et al.* Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. *Genome Res.* **24,** 1854–68 (2014).

76. Jäger, R. *et al.* Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nat. Commun.* **6,** 6178 (2015).

77. Orlando, G. *et al.* Promoter capture Hi-C-based identification of recurrent noncoding mutations in colorectal cancer. *Nat. Genet.* **50,** 1375–1380 (2018).

78. Mumbach, M. R. *et al.* Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. *Nat. Genet.* **49,** 1602–1612 (2017).

79. Burren, O. S. *et al.* Chromosome contacts in activated T cells identify autoimmune disease candidate genes. *Genome Biol.* **18,** 165 (2017).

80. Choy, M.-K. *et al.* Promoter interactome of human embryonic stem cell-derived cardiomyocytes connects GWAS regions to cardiac gene networks. *Nat. Commun.* **9,** 2526 (2018).

81. Khera, A. V. *et al.* Genetic Risk, Adherence to a Healthy Lifestyle, and Coronary Disease. *N. Engl. J. Med.* **375,** 2349–2358 (2016).

82. Musunuru, K. *et al.* From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* **466,** 714–9 (2010).

83. Montefiori, L. E. *et al.* A promoter interaction map for cardiovascular disease genetics. *Elife* **7,** 1–35 (2018).

84. Gu, W. *et al.* Depletion of Abundant Sequences by Hybridization (DASH): using Cas9 to remove unwanted high-abundance species in sequencing libraries and molecular counting applications. *Genome Biol.* **17,** 41 (2016).

85. Duester, G. Knocking Out Enhancers to Enhance Epigenetic Research. *Trends Genet.* **35,** 89 (2019).

86. Gu, W. *et al.* Depletion of Abundant Sequences by Hybridization (DASH): using Cas9 to remove unwanted high-abundance species in sequencing libraries and molecular counting applications. *Genome Biol.* **17,** 41 (2016).

87. Wu, J. *et al.* The landscape of accessible chromatin in mammalian preimplantation embryos. *Nature* **534,** 652–657 (2016).

88. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339,** 819–23 (2013).

89. Jubs. SEQanswers online forum: Mitochondrial contamination. (2014). Available at: http://seqanswers.com/forums/showpost.php?p=133105&postcount=3. (Accessed: 18th January 2017)

90. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell* **38,** 576–589 (2010).

91. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9,** R137 (2008).

92. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22,** 1760–74 (2012).

93. Corces, M. R. *et al.* Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.* **48,** 1193–1203 (2016).

94. Lin, S. *et al.* Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. *Elife* **3,** e04766 (2014).

95. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 10 (2011).

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

97. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–2 (2010).

98. Cooper, G. M. & Brown, C. D. Qualifying the relationship between sequence conservation and molecular function. *Genome Res.* **18,** 201–205 (2008).

99. Sandelin, A. *et al.* Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics* **5,** 99 (2004).

100. Visel, A. *et al.* Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat. Genet.* **40,** 158–160 (2008).

101. Calin, G. A. *et al.* Ultraconserved Regions Encoding ncRNAs Are Altered in Human Leukemias and Carcinomas. *Cancer Cell* **12,** 215–229 (2007).

102. Viturawong, T., Meissner, F., Butter, F. & Mann, M. A DNA-Centric Protein Interaction Map of Ultraconserved Elements Reveals Contribution of Transcription Factor Binding Hubs to Conservation. *Cell Rep.* **5,** 531–545 (2013).

103. De La Calle-Mustienes, E. *et al.* A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate Iroquois cluster gene deserts. *Genome Res.* **15,** 1061–1072 (2005).

104. Peters, T., Dildrop, R., Ausmeier, K. & Rüther, U. Organization of mouse Iroquois homeobox genes in two clusters suggests a conserved regulation and function in vertebrate development. *Genome Res.* **10,** 1453–62 (2000).

105. Cavodeassi, F., Modolell, J. & Gómez-Skarmeta, J. L. The Iroquois family of genes: from body building to neural patterning. *Development* **128,** 2847–2855 (2001).

106. Robertshaw, E., Matsumoto, K., Lumsden, A. & Kiecker, C. Irx3 and Pax6 establish differential competence for Shh-mediated induction of GABAergic and glutamatergic neurons of the thalamus. *Proc. Natl. Acad. Sci. U. S. A.* **110,** E3919-26 (2013).

107. Kim, K. H., Rosen, A., Bruneau, B. G., Hui, C. C. & Backx, P. H. Iroquois homeodomain transcription factors in heart development and function. *Circ. Res.* **110,** 1513–1524 (2012).

108. Becker, M. B., Zulch, A., Bosse, A. & Gruss, P. Irx1 and Irx2 expression in early lung development. *Mech. Dev.* **106,** 155–158 (2001).

109. Marra, A. N. & Wingert, R. A. Roles of Iroquois Transcription Factors in Kidney Development. *Cell Dev. Biol.* **3,** 1000131 (2014).

110. Claussnitzer, M. *et al. FTO* Obesity Variant Circuitry and Adipocyte Browning in Humans. *N. Engl. J. Med.* **373,** 895–907 (2015).

111. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser--a database of tissue-specific human enhancers. *Nucleic Acids Res.* **35,** D88-92 (2007).

112. Tena, J. J. *et al.* An evolutionarily conserved three-dimensional structure in the vertebrate Irx clusters facilitates enhancer sharing and coregulation. *Nat. Commun.* **2,** 310 (2011).

113. Derti, A., Roth, F. P., Church, G. M. & Wu, C. Mammalian ultraconserved elements are strongly depleted among segmental duplications and copy number variants. *Nat. Genet.* **38,** 1216–1220 (2006).

114. McCole, R. B., Fonseka, C. Y., Koren, A. & Wu, C. ting. Abnormal Dosage of Ultraconserved Elements Is Highly Disfavored in Healthy Cells but Not Cancer Cells. *PLoS Genet.* **10,** (2014).

115. Osterwalder, M. *et al.* Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature* **554,** 239–243 (2018).

116. Nord, A. S. *et al.* Rapid and pervasive changes in genome-wide enhancer usage during mammalian development. *Cell* **155,** 1521–1531 (2013).

117. Crawley, J. N. *What's Wrong With My Mouse?: Behavioral Phenotyping of Transgenic and Knockout Mice: Second Edition*. *What's Wrong With My Mouse?: Behavioral Phenotyping of Transgenic and Knockout Mice: Second Edition* (2006). doi:10.1002/9780470119051

118. Reed, D. R., Lawler, M. P. & Tordoff, M. G. Reduced body weight is a common effect of gene knockout in mice. *BMC Genet.* **9,** 4 (2008).

119. Jao, L.-E., Wente, S. R. & Chen, W. Efficient multiplex biallelic zebrafish genome editing using a CRISPR nuclease system. *Proc. Natl. Acad. Sci.* **110,** 13904–13909 (2013).

120. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14,** 417–419 (2017).

121. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research* **4,** 1521 (2015).

122. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15,** 550 (2014).

123. Juan, A. H. *et al.* Enhancer timing of Hox gene expression: deletion of the endogenous Hoxc8 early enhancer. *Development* **130,** 4823–34 (2003).

124. Lam, D. D. *et al.* Partially Redundant Enhancers Cooperatively Maintain Mammalian Pomc Expression Above a Critical Functional Threshold. *PLOS Genet.* **11,** e1004935 (2015).

125. Leighton, P. A., Saam, J. R., Ingram, R. S., Stewart, C. L. & Tilghman, S. M. An enhancer deletion affects both H19 and Igf2 expression. *Genes Dev.* **9,** 2079–89 (1995).

126. Visel, A. *et al.* Targeted deletion of the 9p21 non-coding coronary artery disease risk interval in mice. *Nature* **464,** 409–412 (2010).

127. Sagai, T. *et al.* Elimination of a long-range cis-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb. *Development* **132,** 797–803 (2005).

128. Dickel, D. E. *et al.* Genome-wide compendium and functional assessment of in vivo heart enhancers. *Nat. Commun.* **7,** 12923 (2016).

129. Jeong, Y. *et al.* Spatial and temporal requirements for sonic hedgehog in the regulation of thalamic interneuron identity. *Development* **138,** 531–41 (2011).

130. Mochizuki, Y. *et al.* Combinatorial CRISPR/Cas9 Approach to Elucidate a Far-Upstream Enhancer Complex for Tissue-Specific Sox9 Expression. *Dev. Cell* **46,** 794–806.e6 (2018).

131. Dave, K. *et al.* Mice deficient of Myc super-enhancer region reveal differential control mechanism between normal and pathological growth. *Elife* **6,** (2017).

132. Sur, I. K. *et al.* Mice lacking a Myc enhancer that includes human SNP rs6983267 are resistant to intestinal tumors. *Science* **338,** 1360–3 (2012).

133. Saito, T., Hara, S., Tamano, M., Asahara, H. & Takada, S. *Deletion of conserved sequences in IG-DMR at Dlk1-Gtl2 locus suggests their involvement in expression of paternally expressed genes in mice*. *Original Article-Reprod. Dev* **63,** (2017).

134. Fazel Darbandi, S. *et al.* Functional consequences of I56ii Dlx enhancer deletion in the developing mouse forebrain. *Dev. Biol.* **420,** 32–42 (2016).

135. Bouvier, G. *et al.* Deletion of the mouse T-cell receptor beta gene enhancer blocks alphabeta T-cell development. *Proc. Natl. Acad. Sci. U. S. A.* **93,** 7877–81 (1996).

136. Bahr, C. *et al.* A Myc enhancer cluster regulates normal and leukaemic haematopoietic stem cell hierarchies. *Nature* **553,** 515–520 (2018).

137. Onal, M., St. John, H. C., Danielson, A. L. & Pike, J. W. Deletion of the Distal *Tnfsf11* RL-D2 Enhancer That Contributes to PTH-Mediated RANKL Expression in Osteoblast Lineage Cells Results in a High Bone Mass Phenotype in Mice. *J. Bone Miner. Res.* **31,** 416–429 (2016).

138. Onal, M. *et al.* Unique Distal Enhancers Linked to the Mouse Tnfsf11 Gene Direct Tissue-Specific and Inflammation-Induced Expression of RANKL. *Endocrinology* **157,** 482–96 (2016).

139. Adams, D. J. *et al.* Renin enhancer is critical for control of renin gene expression and cardiovascular function. *J. Biol. Chem.* **281,** 31753–61 (2006).

140. Maurano, M. T. *et al.* Systematic Localization of Common Disease-Associate Variation in Regulatorty DNA. *Science (80-.).* **337,** 1190–1195 (2012).

141. Wright, J. B., Brown, S. J. & Cole, M. D. Upregulation of c-MYC in cis through a Large Chromatin Loop Linked to a Cancer Risk-Associated Single-Nucleotide Polymorphism in Colorectal Cancer Cells. *Mol. Cell. Biol.* **30,** 1411–1420 (2010).

142. Cowper-Sal·lari, R. *et al.* Breast cancer risk–associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nat. Genet.* **44,** 1191–1198 (2012).

143. Rao, S. S. P. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159,** 1665–1680 (2014).

144. Jin, F. *et al.* A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503,** 290–4 (2013).

145. Schmitt, A. D. *et al.* A Compendium of Chromatin Contact Maps Reveals Spatially Active

Regions in the Human Genome. *Cell Rep.* **17,** 2042–2059 (2016).

146. Javierre, B. M. *et al.* Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* **167,** 1369–1384.e19 (2016).

147. Freire-Pritchett, P. *et al.* Global reorganisation of *cis* -regulatory units upon lineage commitment of human embryonic stem cells. *Elife* **6,** e21926 (2017).

148. Rubin, A. J. *et al.* Lineage-specific dynamic and pre-established enhancer–promoter contacts cooperate in terminal differentiation. *Nat. Genet.* (2017). doi:10.1038/ng.3935

149. Siersbæk, R. *et al.* Dynamic Rewiring of Promoter-Anchored Chromatin Loops during Adipocyte Differentiation. *Mol. Cell* **66,** 420–435.e5 (2017).

150. Burridge, P. W. *et al.* Chemically defined generation of human cardiomyocytes. *Nat. Methods* **11,** 855–60 (2014).

151. Cairns, J. *et al.* CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. *Genome Biol.* **17,** 127 (2016).

152. Dao, L. T. M. *et al.* Genome-wide characterization of mammalian promoters with distal enhancer functions. *Nat. Genet.* **49,** 1073–1081 (2017).

153. Diao, Y. *et al.* A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. *Nat. Methods* **14,** 629–635 (2017).

154. Watt, A. J., Battle, M. A., Li, J. & Duncan, S. A. GATA4 is essential for formation of the proepicardium and regulates cardiogenesis. *Proc. Natl. Acad. Sci.* **101,** 12573–12578 (2004).

155. Pikkarainen, S., Tokola, H., Kerkelä, R. & Ruskoaho, H. GATA transcription factors in the developing and adult heart. *Cardiovasc. Res.* **63,** 196–207 (2004).

156. Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A. & Bejerano, G. Enhancers: five essential questions. *Nat. Rev. Genet.* **14,** 288–295 (2013).

157. Miele, A. & Dekker, J. Long-range chromosomal interactions and gene regulation. *Mol. Biosyst.* **4,** 1046 (2008).

158. Deng, W. *et al.* Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell* **149,** 1233–1244 (2012).

159. Phillips, J. E. & Corces, V. G. CTCF: master weaver of the genome. *Cell* **137,** 1194–211 (2009).

160. Phillips-Cremins, J. E. *et al.* Architectural Protein Subclasses Shape 3D Organization of Genomes during Lineage Commitment. *Cell* **153,** 1281–1295 (2013).

161. Nora, E. P. *et al.* Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell* **169,** 930–944.e22 (2017).

162. Cai, C.-L. *et al.* T-box genes coordinate regional rates of proliferation and regional specification during cardiogenesis. *Development* **132,** 2475–2487 (2005).

163. Sakabe, N. J. *et al.* Dual transcriptional activator and repressor roles of TBX20 regulate adult cardiac structure and function. *Hum. Mol. Genet.* **21,** 2194–2204 (2012).

164. Mahmoud, A. I. *et al.* Meis1 regulates postnatal cardiomyocyte cell cycle arrest. *Nature* **497,** 249–253 (2013).

165. Shen, T. *et al.* Tbx20 regulates a genetic program essential to adult mouse cardiomyocyte function. *J. Clin. Invest.* **121,** 4640–54 (2011).

166. Phanstiel, D. H. *et al.* Static and Dynamic DNA Loops form AP-1-Bound Activation Hubs during Macrophage Development. *Mol. Cell* **67,** 1037–1048.e6 (2017).

167. Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459,** 108–12 (2009).

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

169. Erceg, J. *et al.* Dual functionality of cis-regulatory elements as developmental enhancers and Polycomb response elements. *Genes Dev.* **31,** 590–602 (2017).

170. Karlić, R., Chung, H.-R., Lasserre, J., Vlahovicek, K. & Vingron, M. Histone modification levels are predictive for gene expression. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 2926–31 (2010).

171. Stevens, T. J. *et al.* 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature* **544,** 59–64 (2017).

172. Gilbert, N. *et al.* Chromatin Architecture of the Human Genome: Gene-Rich Domains Are

Enriched in Open Chromatin Fibers. *Cell* **118,** 555–566 (2004).

173. Blake, J. A. *et al.* Mouse Genome Database (MGD)-2017: community knowledge resource for the laboratory mouse. *Nucleic Acids Res.* **45,** D723–D729 (2017).

174. Smemo, S. *et al.* Regulatory variation in a TBX5 enhancer leads to isolated congenital heart disease. *Hum. Mol. Genet.* **21,** 3255–63 (2012).

175. Arnolds, D. E. *et al.* TBX5 drives Scn5a expression to regulate cardiac conduction system function. *J. Clin. Invest.* **122,** 2509–2518 (2012).

176. Moshal, K. *et al.* LITAF, A Novel Regulator of Cardiac Excitation. *FASEB J.* **31,** 686.3-686.3 (2017).

177. Arking, D. E. *et al.* Genetic association study of QT interval highlights role for calcium signaling pathways in myocardial repolarization. *Nat. Genet.* **46,** 826–836 (2014).

178. Petersen, C. M. *et al.* Molecular identification of a novel candidate sorting receptor purified from human brain by receptor-associated protein affinity chromatography. *J. Biol. Chem.* **272,** 3599–605 (1997).

179. Smith, N. L. *et al.* Association of Genome-Wide Variation With the Risk of Incident Heart Failure in Adults of European and African Ancestry: A Prospective Meta-Analysis From the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium. *Circ. Cardiovasc. Genet.* **3,** 256–266 (2010).

180. Guo, D.-C. *et al.* Mutations in smooth muscle alpha-actin (ACTA2) cause coronary artery disease, stroke, and Moyamoya disease, along with thoracic aortic disease. *Am. J. Hum. Genet.* **84,** 617–27 (2009).

181. Dekker, J. & Mirny, L. The 3D Genome as Moderator of Chromosomal Communication. *Cell* **164,** 1110–1121 (2016).

182. Calo, E. & Wysocka, J. Modification of enhancer chromatin: what, how, and why? *Mol. Cell* **49,** 825–37 (2013).

183. Hnisz, D., Shrinivas, K., Young, R. A., Chakraborty, A. K. & Sharp, P. A. Perspective A Phase Separation Model for Transcriptional Control. *Cell* **169,** 13–23 (2017).

184. Gherghiceanu, M. *et al.* Cardiomyocytes derived from human embryonic and induced pluripotent stem cells: comparative ultrastructure. *J. Cell. Mol. Med.* **15,** 2539–2551 (2011).

185. Karakikes, I., Ameen, M., Termglinchan, V. & Wu, J. C. Human Induced Pluripotent Stem Cell–Derived Cardiomyocytes. *Circ. Res.* **117,** (2015).

186. Zhou, X. *et al.* Epigenomic annotation of genetic variants using the Roadmap Epigenome Browser. *Nat. Biotechnol.* **33,** 345–346 (2015).

187. Banovich, N. E. *et al.* Impact of regulatory variation across human iPSCs and differentiated cells. *Genome Res.* **28,** 122–131 (2018).

188. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44,** D733-45 (2016).

189. Speir, M. L. *et al.* The UCSC Genome Browser database: 2016 update. *Nucleic Acids Res.* **44,** D717–D725 (2016).

190. Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* **27,** 182–9 (2009).

191. Wingett, S. *et al.* HiCUP: pipeline for mapping and processing Hi-C data. *F1000Research* **4,** 1310 (2015).

192. Shin, H. *et al.* TopDom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Res.* **44,** e70 (2016).

193. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25,** 25–9 (2000).

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

195. Nikpay, M. *et al.* A comprehensive 1000 Genomes–based genome-wide association meta-analysis of coronary artery disease. *Nat. Genet.* **47,** 1121–1130 (2015).

196. Li, H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics* **27,** 718–9 (2011).

197. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–75 (2007).

198. Meder, B. *et al.* A genome-wide association study identifies 6p21 as novel risk locus for dilated cardiomyopathy. *Eur. Heart J.* **35,** 1069–1077 (2014).

199. Carithers, L. J. *et al.* A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. *Biopreserv. Biobank.* **13,** 311–319 (2015).

200. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 9440–5 (2003).

201. Cho, W.-K. *et al.* Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. *Science* **361,** 412–415 (2018).

202. Sabari, B. R. *et al.* Coactivator condensation at super-enhancers links phase separation and gene control. *Science* eaar3958 (2018). doi:10.1126/science.aar3958

203. Chong, S. *et al.* Imaging dynamic and selective low-complexity domain interactions that control gene transcription. *Science* eaar2555 (2018). doi:10.1126/science.aar2555

204. Barolo, S. Shadow enhancers: frequently asked questions about distributed cis-regulatory information and enhancer redundancy. *Bioessays* **34,** 135–41 (2012).

205. Gallagher, M. D. & Chen-Plotkin, A. S. The Post-GWAS Era: From Association to Function. *Am. J. Hum. Genet.* **102,** 717–730 (2018).

206. Starita, L. M. *et al.* Variant Interpretation: Functional Assays to the Rescue. *Am. J. Hum. Genet.* **101,** 315–325 (2017).

207. Nelson, M. R. *et al.* The support of human genetic evidence for approved drug indications. *Nat. Genet.* **47,** 856–860 (2015).

208. King, E. A., Davis, J. W. & Degner, J. F. Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. (2019). doi:10.1101/513945