THE UNIVERSITY OF CHICAGO

SINGLE-CELL GENOMIC APPROACHES FOR INTERPRETING THE GENETIC ARCHITECTURE OF COMPLEX TRAITS

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES

AND

THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES

AND THE PRITZKER SCHOOL OF MEDICINE

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN BIOPHYSICAL SCIENCES

BY

ALAN SELEWA

CHICAGO, ILLINOIS

AUGUST 2021

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like acknowledge all the people who have been patient with me and supported me. In particular my advisors, Professor Xin He and Assistant Professor Anindita "Oni" Basu, have been incredibly patient with me and supported me through-out my PhD. I came from a Physics background and did not have much formal training in genomics prior to my PhD, however Xin and Oni were happy to take me on and teach me the foundations of the field and for that I am very grateful. As a consequence, I always strived to be a fast learner and emulate their expertise. This was particularly challenging in the beginning stages of my PhD where my background was still in development. Having worked on mostly computational projects, I spent a significant amount of time learning the field with Xin. When I didn't know something, Xin would zealously offer multiple resources for learning the same concepts, such as books and literature reviews. If the material still didn't make sense, he would take the time to tutor me so that the concepts solidified. I am grateful for the time he dedicated to his students to help them learn the foundations so that we can go on to do our best work. I'm sure it wasn't easy for him to work with me, especially early on in my PhD when my expertise was lacking and I am grateful for his believing in me. I am also thankful for the personal development that I achieved throughout all my time with Oni. She has cultivated a culture of openness and positivity in the group, and I always felt comfortable to share my circumstances with her, positive or negative. These conversations have led to bettering myself not only in a scientific capacity, but also in softer skills.

I also want to acknowledge Sebastian "Seb" Pott, who has played many roles in my PhD: a scientific collaborator, a consultant, and as a mentor. I owe much of the data I analyzed to the experiments that he designed and piloted, and I am grateful to have him as a scientific collaborator. Seb advised me on many aspects of my projects and working with him has made me a better

scientist. Because of him, I learned the experimental details much more intimately, which allowed me to be a much more effective computational scientist. He has always made time available to listen to my ideas and answer my questions about science, technology, careers, and for that I am grateful.

I want to express my gratitude for my committee members, Matthew Stephens and Megan McNerney. Megan has given me excellent feedback for all my projects, especially the cancer-related projects. I am always impressed and appreciative of her inquisitiveness, particularly on the biological details that I am sometimes miss. Being in a meeting with Matthew, whether he is a presenter or audience member, usually leads to me learning something new and insightful. Matthew has an incredible ability to get to the essence of a scientific problem and re-frame it in a clear and simple way. I believe my projects would not have gone as far as they did without the contributions of Matthew and Megan, and I am thankful for that.

Finally, I want to acknowledge the people nearest and dearest to me who have shaped who I am today. I first want to thank my father, who did not attend higher education, but has always recognized the importance of education. He worked hard to emigrate our family to the United States where I can have access to better education. He was always supportive of my pursuing the PhD and encouraged me to continue in difficult times, so for that I thank him. I also want to acknowledge my girlfriend Yifan Zhou, who has been incredibly supportive throughout this PhD. I am eternally grateful for all the scientific discussions we had and all the emotional support she has provided.

ABSTRACT

Variation in DNA sequence influences change in one or many molecular intermediates in a functional pathway, ultimately leading to a change in an organismal-level trait. This creates a causal chain of events, as governed by the Central Dogma of molecular biology, where deleterious DNA variants cause dysregulation of gene expression and/or protein levels, leading to a disease state at the organismal-level. Determining which and how DNA variants are causal for the disease phenotype is a major challenge in the field of genetics and is of major interest due to its potential for unraveling new knowledge about regulatory biology and discovering new genetic therapies for diseases. Single nucleotide variants (SNVs), or just variants, can be classified into two classes: namely single-nucleotide polymorphism (SNPs) which occur at some frequency in the human population, and somatic point mutations which occur throughout the lifespan of the organism. The vast majority of disease-associated variants tend to be in the non-coding part of the genome, leading to complex and variable interactions with genes. Perhaps the best understood of these non-coding variants are regulatory variants which reside in DNA regulatory elements such as promoters, enhancers and repressors. The activity of regulatory elements has been shown to be cell-type and state specific, which motivates the need for single-cell technologies for further dissecting disease-related variants and the putative genes they target. In this dissertation, I develop a framework for utilizing single-cell 'omics data to interpret the germline SNPs and somatic point mutations associated with disease states. In Chapter 1, I explore methods for detecting somatic mutations in individual cancer cells and nominate genes whose expression is altered in cells with somatic mutations using single-cell RNA-sequencing data. However, obtaining single-cell RNA-sequencing data from bulk tissues such as solid tumors presents its own challenges. Due to the complexity of the intracellular matrix of adult bulk tissues, such as solid tumors, obtaining single

cell suspensions is not always possible. In Chapter 2, I performed a systematic analysis between single-cell and nucleus RNA-sequencing data on a model system of induced-pluripotent stem cells differentiating into cardiomyocytes. Finally, I developed a framework in Chapter 3 for utilizing single-nucleus ATAC-seq and single-nucleus RNA-seq to interpret the germline SNPs found in atrial fibrillation (AF) GWAS, the most common cardiac arrhythmia. Risk variants of Atrial Fibrillation (AF) are >10-fold enriched in cardiomyocytes (CMs) but not other cell types. Taking advantage of this enrichment pattern, we used a Bayesian statistical framework to fine-map causal variants of AF, favoring variants in CM open chromatin regions. I developed a novel computational procedure that aggregates all putative causal variants and combines multiple sources of information linking SNPs to genes. Through this procedure, I nominate genes that were not found by GWAS alone.

INTRODUCTION

Variation in DNA sequence influences change in one or many molecular intermediates in a functional pathway[1], ultimately leading to a change in an organismal-level trait. This creates a causal chain of events, as governed by the Central Dogma of molecular biology, where deleterious DNA variants cause dysregulation of gene expression and/or protein levels, leading to a disease state at the organismal-level. Determining which and how DNA variants are causal for the disease phenotype is a major challenge in the field of genetics and is of major interest due to its potential for unraveling new knowledge about regulatory biology and discovering new genetic therapies for diseases.

Single nucleotide variants (SNVs), or just variants, can be classified into two classes: namely single-nucleotide polymorphism (SNPs) which occur at some frequency in the human population, and somatic point mutations which occur throughout the lifespan of the organism and are generally unique to that organism. Somatic mutational burden is known to partly drive cancer and tumorigenesis by mutating key genes such as tumor suppressor genes and oncogenes[2]. On a larger scale, genome-wide association studies (GWAS) have found countless SNPs genome-wide that are significantly associated with thousands of traits[3] through rigorous statistical testing and functional validation.

In the route of assessing which and how DNA variants influence a certain trait, e.g. cancer, the first link in the causal chain is inferring the variant and the gene it influences. A variant, whether it is a SNP or somatic, can be categorized as coding or non-coding depending on its genomic position. In the former case, the functional consequence of the variant can be inferred simply by its influence on the primary amino acid structure. For instance, a non-sense variant in the start

codon of a gene can completely shut-off expression. Such mutations are known as loss-of-function (LoF) mutations and frequently occur in cancer by disrupting expression of key tumor suppressor genes such as TP53.

The vast majority of disease-associated variants tend to be in the non-coding part of the genome, leading to complex and variable interactions with genes. Perhaps the best understood of these non-coding variants are regulatory variants which reside in DNA regulatory elements such as promoters, enhancers and repressors. The activity of regulatory elements has been shown to be cell-type and state specific[4], which motivates the need for single-cell technologies for further dissecting disease-related variants and the putative genes they target.

Recent advances in molecular biology, microfluidics and nanotechnology have given rise to a multitude of single-cell sequencing technologies. Initial methods have focused on measurements of a single molecular entity such as DNA sequence, transcriptomics, and chromatin accessibility. At all levels of the central dogma hierarchy, single-cell measurements have revealed extensive cellular heterogeneity in the composition of complex tissues in a variety of organisms such as mouse retina[5], malaria parasites[6], and drosophila embryos[7]. A notable application of single-cell technologies is in the field of cancer biology where single-cell transcriptomics were used to dissect intratumor cellular heterogeneity and establish the tumor microenvironment[8]. Single-cell multimodal assays combining DNA sequence and transcriptome readout provide a powerful technique for inferring the marginal effect of a SNV on the entire transcriptome directly[9], [10]. Single-cell assays that probe epigenetics such as chromatin accessibility allow us to infer regulatory networks, which tend to be cell-type and tissue specific. Furthermore, variants falling into regulatory elements can potentially be linked to their putative target gene in a cell-type-

specific manner. Overall, single-cell ‘omics provide cellular-resolved functional genomics for dissecting effects of DNA variants on gene expression.

In this dissertation, I develop a framework for utilizing single-cell ‘omics data to interpret the germline SNPs and somatic point mutations associated with disease states. This interpretation involves learning the relevant cell-type(s) for the disease, and the cell-type specific intermediate genes that mediate the effect of the SNP onto the phenotype. Here we refer to cell types/states in the most generic sense; they can be defined by genetic, epigenetic, and/or transcriptional profiles. In Chapter 1, I explore methods for detecting somatic mutations in individual cancer cells and clustered cells according to their somatic mutational profiles. In particular, I develop a pipeline for genotyping individual cells using full-transcript single-cell RNA-sequencing data, and perform single-cell phylogenetic inference based on somatic mutation status in each cell. We utilized a recently published method called Single Cell Inference of Tumor Evolution, SCITE[11], which allows us to effectively cluster single cancer cells based on their evolutionary history as implied by somatic mutations. Once we defined groups of evolutionary-similar cells, we perform traditional differential expression analysis across cell clusters to find genes potentially impacted by somatic point mutations, and to infer how gene expression changes during the tumor evolution.

However, obtaining single-cell RNA-sequencing data from bulk tissues such as solid tumors presents its own challenges. Due to the complexity of the intracellular matrix of adult bulk tissues, such as solid tumors, obtaining single cell suspensions is not always possible. An alternative approach is to use single nuclei[12], however this potentially problematic for single-cell RNA-sequencing because there is less RNA content to sequence. In Chapter 2, I perform a systematic analysis between single-cell and nucleus RNA-sequencing data on a model system of induced-

pluripotent stem cells differentiating into cardiomyocytes. I utilized state-of-the-art methods for batch correction, dimensionality reduction, clustering, and differentiation trajectory reconstruction on both modalities. I mapped and compared the cell-types and states found between the two modalities to assess their potential for detecting common and rare cell-types. I also assessed the difference in average genome-wide coverage, a critical parameter for the genotyping algorithm discussed in Chapter 1.

In Chapter 3, I developed a framework for utilizing single-nucleus ATAC-seq to interpret the germline variants found in atrial fibrillation (AF) GWAS. AF is the most common cardiac arrhythmia and affects 2-9% of people worldwide and varies with age. Recent large-scale genome-wide associations have discovered over 111 independent loci that contribute to AF risk[13] corresponding to 122 approximately independent linkage disequilibrium (LD) blocks; however the causal variants and their target genes remain unknown. Utilizing our cell-type resolved cardiac single-cell ATAC-seq data, we assessed the enrichment of cell-type specific open chromatin regions (OCRs) in AF GWAS SNPs. I found up to 8-fold enrichment of cardiomyocyte-specific OCRs in AF, while very little enrichment was observed in other cell-types of the heart. I took advantage of this strong enrichment by performing Bayesian statistical fine-mapping using SuSiE[14] in order to prioritize AF SNPs in cardiomyocyte OCRs. Using this strategy, we were able to identify new AF-associated variants that would not have been identified without the functional information. Utilizing promoter-capture HiC (PC) and coaccessibility between distal enhancers and promoter regulatory elements, I linked the fine-mapped causal variants to their putative target genes. The genes we found were differentially expressed and upregulated in cardiomyocytes compared with other cell-types of the heart.

CHAPTER 1: SINGLE-CELL PHYLOGENETIC INFERENCE FROM SINGLE-CELL RNA AND WHOLE-EXOME SEQUENCING DATA

1.1 Introduction

Tumorigenesis, or the formation of a cancer tumor, has increasingly been accepted as an evolutionary process where cell populations evolve much like populations of organisms do[15]. However, unlike large-scale divergence such as between human and chimps, the evolutionary distance between individual cancer cells is extremely short due to the very small divergence in DNA sequence. To set this in perspective, the divergence, as measured by the number of nucleotide substitutions per site, is $\sim 10^{-1}$ between mammals, $\sim 10^{-3}$ between two humans, and about $\sim 10^{-6}$ between two cells[15]. During cancer evolution, many forces operate such as natural selection to produce the cancer cells with most fitness, here defined as differential cellular proliferation among distinct cellular somatic mutations. The consequences of this evolutionary model are groups of tumor cells, or sub-clones, each harboring a set of distinct and advantageous somatic mutations. In the simplest model, also known as the clonal expansion model, these sub-clones compete against each other for resources in the tumor environment and the more fit ones will replace others until eventually they themselves are out-competed by new and more fit sub-clones[16], [17].

The clonal expansion model predicts a high degree of genetic diversity within tumors, and this is referred to as intra-tumor heterogeneity (ITH). ITH is believed to a major cause of cancer relapse after treatment because the drug therapy typically targets the dominant (with highest number of cells) sub-clone present at the time of diagnosis[18], [19]. In an evolutionary framework, the success of this drug therapy implies the removal of one competitor sub-clone. Upon remission, either a clonal expansion of a previously suppressed sub-clone or an emergence of a new resistant

subclone is likely to occur[20]. Knowledge of the genetic diversity and evolutionary history of a tumor is likely to be a key component in the design of personalized cancer therapies that are more effective.

It is important to define the set of sub-clones making up a tumor, e.g. ITH, because they potentially have different molecular phenotypes, e.g. transcriptome, which would be critical to understand prior to designing a drug. The main challenge in obtaining knowledge on ITH, e.g. the set of sub-clones and their distinct somatic mutational signatures, is that common bulk high-throughput sequencing admixes the DNA of millions of cells in a sample before sequencing. The resulting mutational profiles obtained from the mixture constitute an average of an unknown number of sub-clones each making up an unknown fraction of the mixture[21]. An abundance of tools have been developed to infer the sub-clones and their evolutionary history from bulk DNA sequencing data[22]–[24], however their resolution is inherently limited and low-frequency sub-clones are difficult to detect.

The emergence of single-cell DNA sequencing has changed the situation by unlocking the ability to reconstruct single-cell phylogenies. All cells in a tumor are related via a binary genealogical tree. To reconstruct the evolutionary history based on SNVs, or somatic point mutations, the infinite sites assumption is typically made, which implies that the mutation profiles of the cells form a perfect phylogeny. A perfect phylogeny is defined if for all pairs of mutations i_i, i_2, the set of cells having mutation i_i and the set of cells having mutation i_2 are either disjoint or one is a subset of the other. Most approaches constructing tumor phylogenies focus on the order of SNV or mutation events, also known as the mutation tree. The sub-clones, or set of cells harboring distinct mutations, are implicitly inferred by taking all mutations in a path from the root node to any leaf node in the mutation tree. A recent method SCITE[11] uses this scheme to infer mutation

trees, and therefore putative sub-clones, from single-cell DNA sequencing data. SCITE uses an MCMC algorithm to infer the tree that maximizes the posterior probability given the single-cell genotype data.

The inference of sub-clones in tumors is a critical step for defining the set of heterogeneous mutational profiles, however, the downstream effect of these mutations, and consequently the molecular phenotype, are still largely unknown. As we discussed in the introduction section, linking point mutations to their putative target genes is the first link in the causal chain going from DNA variation to phenotype. Using single-cell DNA-sequencing data allows us to gain cellular resolved mutational profiles. Other functional genomics data such as RNA-seq from the same cells would allow us to potentially measure the effects of the somatic mutations. Indeed, a multi-omic single-cell approach combining DNA and RNA simultaneous measurements would be ideal for reconstructing phylogenies and measuring transcriptome differences between sub-clones. However, methods for simultaneous measurement of DNA and RNA are fairly immature, expensive, and low-throughput.

In this chapter, I develop a method that adopts SCITE for inferring tumor sub-clones from single-cell RNA-sequencing combined with a matched bulk-tissue whole-exome sequencing (WES) data. In this way, we circumvent the need for simultaneous measurements of RNA and DNA in individual cells. The advantage of single-cell RNA-sequencing is that we have transcriptome readout for each inferred sub-clone, which allows us to infer changes in molecular intermediates, genes in this case, during the tumor evolution. Given that single-cell RNA-sequencing data are relatively sparse and reads only span gene bodies, genotyping each cell is a challenge, and therefore the quality of reconstructed phylogenies will largely depend on experimental and sequencing parameters of the single-cell RNA-sequencing data. To that end, we performed

extensive simulations using ground-truth phylogenies to assess the performance of the method. Finally, I apply this to publicly available data from a recent work[25] that performed single-cell RNA-sequencing and bulk DNA sequencing from human breast cancer tissue. We find that sub-clones with more somatic mutations show an increased expression of oncogenes and lowered expression of tumor suppressor genes.

1.2 Methods

1.2.1 Computational procedure

In this section, we outline the main pipeline for genotyping individual cells using single-cell RNA-sequencing data (scRNA-seq) paired with matched bulk WES data. The vast majority of somatic mutations are background mutations which are randomly distributed in the tumor genome. However, due to selection the nonsynonymous mutations in exonic regions has increased mutation rate, therefore making RNA-seq potentially suitable candidate for detecting such variants. Calling somatic variants from RNA-seq *de-novo* is, however, a big challenge and previous studies have found limited success with large false positive rates owing to alignment errors, strand bias, cycle bias, random errors during reverse transcription and PCR, and mechanisms such as mRNA-editing[26]. In the case of scRNA-seq data, the above challenges are compounded by low sequencing depth. Here we take an alternative approach and attempt to solve an easier question: given a set of somatic point mutations that are known *a priori*, can we confidently determine whether a cell has the mutation or not based on scRNA-seq data? Utilizing WES data at 100X coverage from the same individual/tumor, we use a GATK-based[27] short somatic variant discovery pipeline to determine the set of somatic point mutations and classify their presence/absence in individual cells using scRNA-seq (Figure 1.1).

Single-cell RNA-sequencing methods can be classified into two categories: those based on droplet 3-prime technology and plate-based technology that sequence full transcripts as opposed to just the 3-prime end. The former is lower cost and higher throughput in terms of number of cells compared with the latter, however, the number of reads per cell are about 100 times smaller on average. Less reads per cell leads to a smaller average genome-wide coverage which makes the 3-prime scRNA-seq data particularly challenging for genotyping individual cells. However, the large number of cells sequenced using 3-prime technologies is an advantage for adequately sampling the diversity of sub-clones in a tumor. We consider both kinds of data in our simulation study of reconstructing phylogenies, and we apply the method to plate-based scRNA-seq data.

The pipeline begins by taking in FASTQ files for tumor scRNA-seq and WES data, and optionally FASTQ files for WES from patient-matched non-tumor (normal) tissue for classifying germline variants. A number of bioinformatics solutions exist for processing whole exome/genome sequencing data. I developed a Snakemake workflow based on GATK's best practices and it can be found at github.com/aselewa/cancerWGS_pipeline. The pipeline utilizes bwa-mem for aligning DNA fragments, followed by de-duplication of PCR duplicates, base-quality score re-calibration, and Mutect2 for short variant discovery. Mutect2 can be run in tumor-only mode, but in this project, we used a matched normal tissue and defined germline variants using HaplotypeCaller. Furthermore, a germline variant resource based on Gnomad is used to further classify somatic vs germline variants. Mutect2 produces a germline posterior probability, which we threshold at 50% and define a set of high-quality short somatic variant calls. For scRNA-seq data, we can utilize another homemade solution available at github.com/aselewa/dropseqRunner for ingesting either 3-prime or full-length scRNA-seq FASTQ files and aligning them to produce BAM files.

Given a set of positions containing somatic point mutations (based on WES), we use a simple model to classify the presence/absence of each mutation in each cell in scRNA-seq BAM files. Using samtools mpileup, we obtain the counts of the reference and mutant allele at all somatic mutation locations in the scRNA-seq BAM file. We retain mutations that have a coverage of at least 10 reads in each cell. Let D_{ij} be the total read count of somatic mutation i in cell j. Using the read count data, we would like to estimate $P(R|G)$ for each genotype G using the read data R and choose the genotype with the highest likelihood. Assuming the reads are independent, the genotype likelihood is given by:

$$P(R|G) = \prod_{k=1}^{D_{ij}} P(r_k|G)$$

where r_k is the kth read and the probability of the read k given the genotype $G=H_1H_2$ is given by:

$$P(r|G) = \frac{1}{2}[P(r_k|H_1)+P(r_k|H_2)]$$

In other words, we average the conditional probability of an observed allele given the genotype alleles. In the simplest model, assuming uniform errors, this condition probability is given by:

$$P(r_k|B) = \begin{cases} 1-\epsilon & D_k=B \\ \frac{\epsilon}{3}, & \text{otherwise} \end{cases}$$

where $\epsilon/3$ is the per-allele sequencing error rate and we set $\epsilon = 10^{-2}$. For downstream analyses and tree construction, we do not distinguish between the cases where a cell contains one or two copies of the somatic mutation allele. We simply denote a binary status of whether the cell contains the somatic mutation or not. The above procedure yields a binary single-cell genotype matrix which

is then passed to SCITE for inferring the mutation tree and subsequently the sub-clones. Missing values in the genotype matrix, due to insufficient read depth in a sub-set of cells or ambiguous genotyping, are allowed as input for SCITE, which reports accurate tree reconstruction with up to 60% missing values in the genotype matrix.

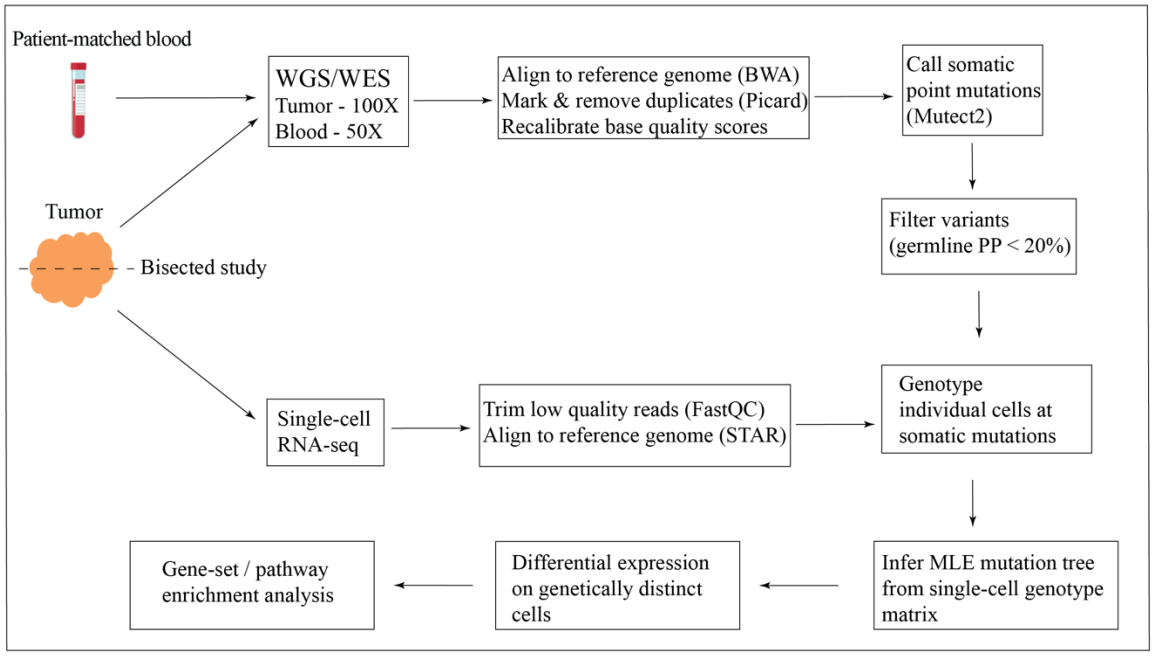

Figure 1.1: Schematic of computational procedure for inferring phylogenies from single-cell RNA-sequencing data. WGS: whole genome sequencing, WES: whole-exome sequencing. PP: posterior probability.

1.2.2 Benchmarking simulated trees at various sparsity levels in scRNA-seq

In this section, we assessed the performance of this procedure by performing simple simulations. We expect that the performance of this procedure relies largely on two parameters: the number of somatic mutations and the number of reads per cell in scRNA-seq. The performance also relies on the number of cells, however, plate-based scRNA-seq experiments are generally low through-put and typical experiments produce on the order of $\sim 10^2$ cells so we fix the number of cells in our simulations. The number of somatic mutations is also determined by the genome-wide mutation rate for the specific cancer type[2], but here we use the mutation rate per Mb averaged over all

cancer types to keep the methodology agnostic to cancer type. Using a mutation rate of 2 per Mb[2], and assuming the exome is 30Mb large, we expect about 60 somatic point mutations.

Constructing a ground truth phylogeny

We begin by constructing a ground truth phylogeny of individual cells belonging to three sub-clones (Figure 1.2). The phylogeny is simply a binary tree of three sub-clones, and the cells within each sub-clone form a star shaped phylogeny. That is to say the cells within a sub-clone have identical mutational profiles. We used 200 cells in the simulation, with sub-clone A, B, and C having 100, 50, and 50 cells, respectively. The number of mutations was fixed to 60, and was distributed among the branches of the phylogeny as shown in Figure 1.2. Based on this ground truth phylogeny, we assemble the genotype matrix G_{ij} that defines the phylogeny (Figure 1.2).

Simulating RNA-seq coverage

To simulate RNA-seq coverage at any given position in a gene, we begin by simulating counts at all genes, and then we distribute the counts among the positions of all genes. Let S be the total number of reads that we will distribute over P genes. To do this, we use a simple multinomial model:

$$R_j \sim \text{Multinom}(S, \alpha_1 \ldots \alpha_P)$$

Where R_j is the coverage profile of the j-th cell across P genes. The parameters $\alpha_1 \ldots \alpha_p$ are the normalized mean expression levels for P genes. A simple approach would be to distribute them evenly, however, we wanted to capture the library complexity of realistic datasets. To determine these, we fit a negative-binomial model to several single-cell RNA-seq datasets containing ~23k genes to estimate the means and proceeded to normalize them such that $\sum_j \alpha_j = 1$.

Next, we choose at random 60 genes to mutate, therefore we assume exactly one mutation per gene. To determine coverage at the positions within these genes, we simply take the read counts

within each gene and distribute them over the length of the gene uniformly. To be more precise, we use the following:

$$D_{ij} = \frac{R_{ij}}{L_j} r$$

Where D_{ij} is the read coverage at the i-th cell, and j-th mutation (or gene), R_{ik} is the number of reads at the i-th cell and j-th gene, L_j is the length of the current gene j, and r is the read length. In other words, we convert the read count into base-pairs using the read length, and then we estimate the average coverage in a gene.

Simulating allele counts at variable loci and performing tree reconstruction

Now that we have generated the read count at each cell and each mutation, we simulate the number of mutant/non-reference alleles based on the ground truth genotype matrix G_{ij} as follows assuming a diploid system:

$$X_{ij} \sim \text{Binom}(D_{ij}, p)$$

$$p = \left(\frac{1}{2}\right)^{G_{ij}} \left(\frac{\epsilon}{3}\right)^{1-G_{ij}}$$

where $G_{ij} \in \{0,1\}$, and the probability p is approximately 0.5 when the mutation is present ($G_{ij} = 1$, heterozygous cell) and p is $\epsilon/3$ when the mutation is absent. Note that this simulation ignores the case when the cell is non-reference homozygous for simplicity. Having generated the read counts, we now perform genotyping on each cell/mutation using D_{ij} and X_{ij} as prescribed in section 1.2.1, which generates a reconstructed genotype matrix G'_{ij}. The goal is now to reconstruct the phylogeny using SCITE and compare with the ground truth. To this end, we also developed a metric for comparing two binary phylogenies that takes sub-clone structure into account. In particular, we would like a metric that describes how well grouped cells from the same sub-clones

are. For each sub-clone, we find the sub-tree of the reconstructed tree that contains the majority of the cells in said sub-clone and we calculate the proportion of cells in that sub-tree that are not in the sub-clone. We repeat this procedure for all sub-clones and estimate the error as follows:

$$W = \frac{1}{N} \sum_{k \in \text{subclones}} n^{k}_{\text{mismatch}}$$

where N is the total number of cells and n^{k}_{mismatch} is the number of cells that do not belong to *k*th sub-clone but localize in the same sub-tree as the *k*th sub-clone. In other words, $\frac{n^{k}_{\text{mismatch}}}{N}$ is the error rate coming from the *k*th sub-clone. Finally, I take our performance metric to be 1-W, with larger values corresponding to trees that are better reconstructed.

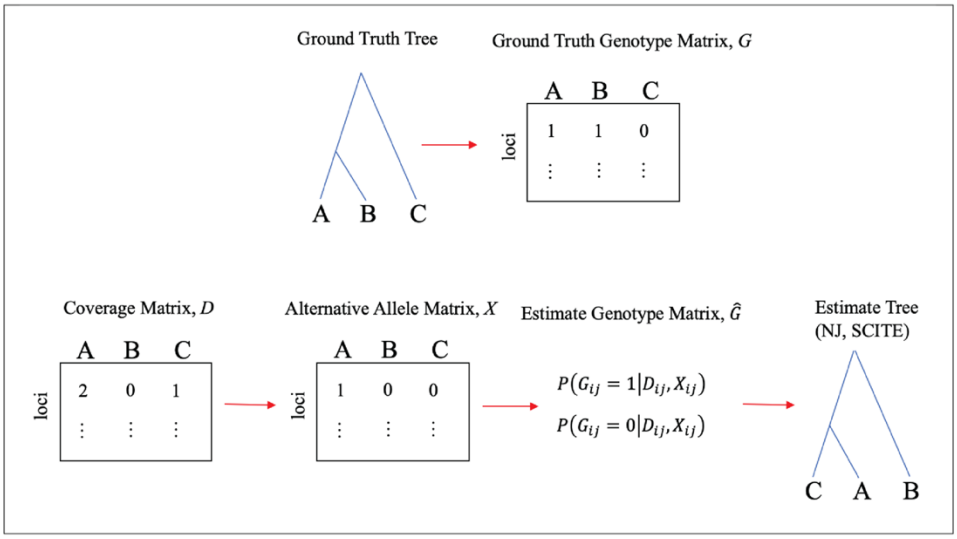

Figure 1.2: Schematic illustrating steps taking to perform benchmarking of phylogenetic inference from single-cell RNA-sequencing data. NJ: neighbor-joining. SCITE: Single cell inference of tumor evolution.

1.3 Results

1.3.1 Simulation results

The above simulation was performed at a number of reads per cell S spanning from 10^4-10^6, which translates to per-base exome coverage values of 0.05 – 10 assuming 100 bp paired-end reads

(Figure 1.3). As a comparison, we used a distance-based neighbor-joining method to reconstruct the phylogeny in addition to SCITE, a maximum-likelihood based method. We find that the performance of both methods generally increases with increasing number of reads per cell, which is likely due to increased probability of having non-zero read count at mutations that segregate and separate cells. Interestingly, the neighbor-joining method shows worst performance at low per-base coverage values compared with SCITE, likely due to mis-calibrated distances between cells because of missing genotypes. Indeed, maximum-likelihood methods such as SCITE have previously been shown to outperform distance-based methods[11]. Based on these simulations, we estimate a minimum single-cell per-base coverage of at least 6X (or 10^6 reads/cell) and ~60 mutations in order to obtain >95% accuracy of reconstruction (Figure 1.3). To reiterate, this level of coverage is common from plate-based full-transcript single-cell transcriptomics. For the remainder of the chapter, we apply the pipeline to previously published plate-based scRNA-seq data with tumor-matched WES data.

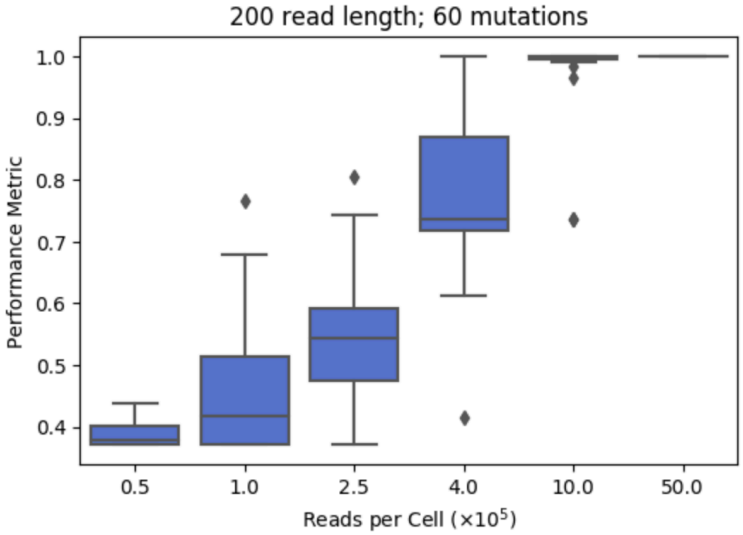

Figure 1.3: Performance of tree reconstruction using SCITE as a function of number of reads per cell with 60 somatic mutations detected from whole exome sequencing data.

1.3.2 Preprocessing scRNA-seq and WES data from breast cancer

A recent study[25] performed comprehensive transcriptomic and genomic profiling of the four subtypes of breast cancer from 11 patients: luminal A/B, HER2, and triple negative. The primary goal of this study was to generate transcriptomic signatures of different cancer sub-types, however, we sought to use the data presented here to demonstrate the applicability of the pipeline. For each patient, tumor biopsies were used to performed scRNA-seq and WES in a bisected experimental design. Peripheral blood was also obtained from each sample to serve as a genomic control for estimating somatic point mutations. The full details of data processing for scRNA-seq and WES can be found in the extended methods of the study. In brief, RNA-seq 100bp paired-end reads were aligned to the human reference genome version 19 (hg19) using the 2-pass mode of STAR[28] version 2.4.0 to produce single-cell BAM files. Exome-sequencing 100bp paired-end reads were aligned to hg19 using BWA[29] version 0.7.1. Putative duplicates were marked by Picard version 1.93. The "best practices" guide for calling short somatic variants from GATK was used which utilizes Mutect2 to call point somatic mutations. The coverage of the WES data was 100x for tumors and 50x for blood samples.

The sequencing-reads quality, number of cells, and number of somatic point mutations varied considerably from one patient to another. Another varying quality was the transcriptomic heterogeneity as demonstrated by PCA in Figure 1a in the referenced study[25]. Indeed, the best sample is one with a large number of cells and somatic point mutations, and exhibits transcriptional heterogeneity which potentially demonstrates multiple sub-clones and/or cell states. In order to demonstrate the pipeline on the best quality dataset we choose one patient that was ER/HER2 positive, namely "BC03", which had a total of 33 cells each with an RNA coverage of at least 10^6 reads/cell, and a total of 429 somatic point mutations called. Approximately 52% (223/429) of

somatic mutations were intronic or intergenic, while 28% (120/429) were missense mutations and 10% (42/429) were silent mutations. Multiple canonical driver genes were found to be mutated, such as proto-oncogene BRAF, HUWE1, and tumor suppressor DEDD2.

1.3.3 Differential expression across two sub-clones

Using the approach outlined in Section 1.2.1, each cell as genotyped at each locus, and we retained 54% of cells (18/33) that had at least 1 somatic mutation. A total of 18 mutations were retained that had at least 10x coverage in one cell (see Table 1.1 for details). We estimate only 5 somatic mutations to be part of driver genes, which indicates that the other 13 mutations are possibly passenger mutations. Nonetheless, passenger mutations are still potentially informative for estimating the evolutionary history of a tumor. We find 61% of somatic mutations (11/18) are missense mutations, while the remaining are intronic, silent, and/or nonsense mutations. A binary single-cell mutation matrix was produced with 66% missing values and a maximum-likelihood mutation tree was inferred with SCITE (Figure 1.4). A total of 100 mutation trees were sampled from the posterior distribution, however, the differences between these trees were local. Single cells were attached to the mutation tree at the most recent mutation harbored. A single-cell may have multiple attachment points due to uncertainty introduced by missing data (Figure 1.4, dashed red lines). Nonetheless, the mutational profile of a single-cell can be inferred by tracing the tree path from the attachment point to the root

Symbol	Chr	hg19 bp	Ref	Alt	Variant_Classification	Genome_Change	Protein_Change	VAF
CRYZ	1	75190429	G	A	Missense_Mutation	g.chr1:75190429G>A	p.S26L	0.04402516
DEDD	1	161094097	G	A	Silent	g.chr1:161094097G>A	p.V52V	0.04375
HSPA6	1	161495788	G	T	Missense_Mutation	g.chr1:161495788G>T	p.G447V	0.0617284
TPR	1	186342575	C	T	Missense_Mutation	g.chr1:186342575C>T	p.E58K	0.04054054
FAM98A	2	33813485	T	C	Missense_Mutation	g.chr2:33813485T>C	p.K147E	0.03589744
PDS5A	4	39924289	G	C	Missense_Mutation	g.chr4:39924289G>C	p.Q203E	0.06818182
CAGE1	6	7339654	C	T	Intron	g.chr6:7339654C>T		0.075
ATG5	6	106740937	G	A	Missense_Mutation	g.chr6:106740937G>A	p.S94L	0.08860759
EFR3A	8	132988310	C	G	Nonsense_Mutation	g.chr8:132988310C>G	p.S399*	0.08571429
RPP30	10	92660338	A	G	Missense_Mutation	g.chr10:92660338A>G	p.T237A	0.046875
ILK	11	6631711	C	T	Missense_Mutation	g.chr11:6631711C>T	p.R410W	0.06306306
APLP2	11	130010337	G	C	Missense_Mutation	g.chr11:130010337G>C	p.E640Q	0.06818182
HEATR5A	14	31872155	C	T	Silent	g.chr14:31872155C>T	p.L9L	0.03249097
EIF5	14	103806854	G	A	Missense_Mutation	g.chr14:103806854G>A	p.E388K	0.03875969
TMEM219	16	29982865	G	A	Silent	g.chr16:29982865G>A	p.*241*	0.0617284
CYTH1	17	76705760	C	T	Missense_Mutation	g.chr17:76705760C>T	p.R26K	0.05147059
SLC25A17	22	41166839	C	T	Silent	g.chr22:41166839C>T	p.*308*	0.1
HUWE1	X	53560335	C	G	Missense_Mutation	g.chrX:53560335C>G	p.E4354Q	0.06818182

Table 1.1: Details of the 18 somatic mutations that are part of the mutation inferred from SCITE.

Next, we separated the cells into two categories in order to variable gene expression across the phylogeny: early and late cells. The late cells generally at least 3-4 more genes mutated than early cells. Because early cells precede the late cells in the evolutionary history of the tumor, we can consider these as two sub-clones. We first assessed the oncogene (ERBB2/3, PIK3CA, CCND1) and tumor suppressor gene (BRCA1/2, ATM, TP53) expression between the early and late sub-clones (Figure 1.5A). We generally find increased oncogene expression and decreased tumor suppressor gene expression in late vs. early sub-clones. In particular, we find 3X more CCND1 expression in late vs. early, while BRCA1/2 expression was about half in late vs. early in terms of cell-averaged transcripts per million (TPM).

The advantage of this approach is the ability to perform differential expression (DE) across the evolutionary history of the tumor in agnostic fashion. While this approach does not explicitly link DE genes to somatic mutations, it highlights changes in the transcriptome during the tumor evolution. To this end, we performed DE across the late vs early cells using a t-test. At FDR < 10%

and a $\log_2$-fold change greater than 1 in the late cells, we found 42 genes that are up-regulated in the late cells. Despite 11 genes having missense mutations in our phylogeny, none of these genes are strongly differentially upregulated in the late cells possibly due to nonfunctional amino acid substitutions. However, we observe strong up-regulation of previously noted oncogenes, as well as RAB3D which has been reported to promote breast cancer cell invasion and metastasis[30]. Using these 42 genes, we performed gene-set enrichment analysis (GSEA) and found 12 GO terms with FDR < 5%. The most enriched categories were intermediate filaments, ATPase complex, and plasma membrane related molecular annotations (Figure 1.5B).

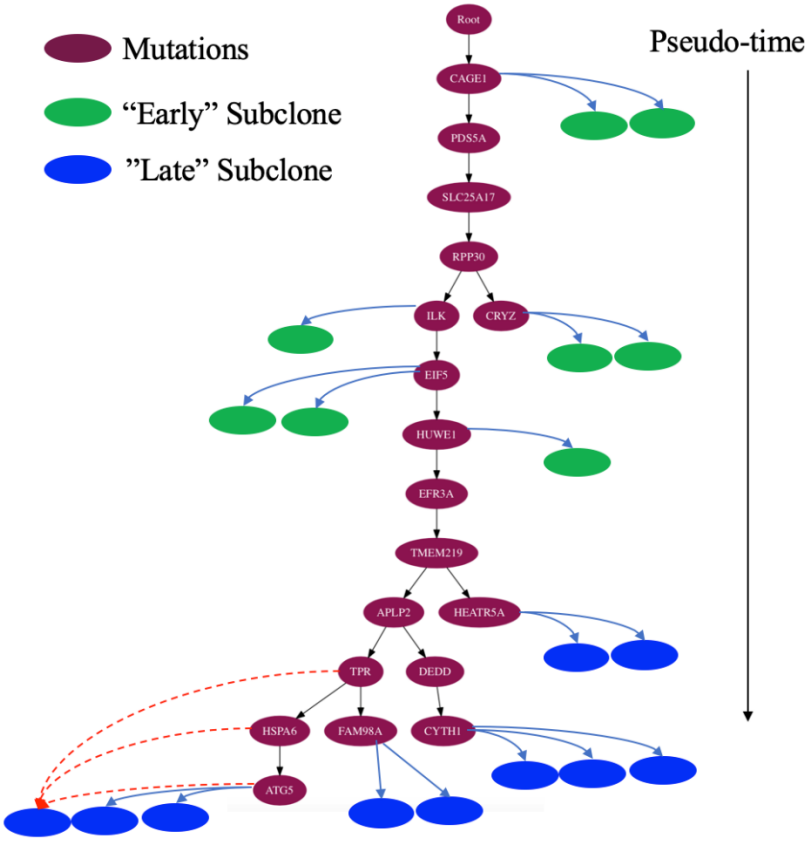

Figure 1.4: Maximum-likelihood mutation tree inferred from SCITE. Dark red nodes are mutations, and the green and blue nodes are single cells. Dashed red lines highlight a certain that has an uncertain attachment point.

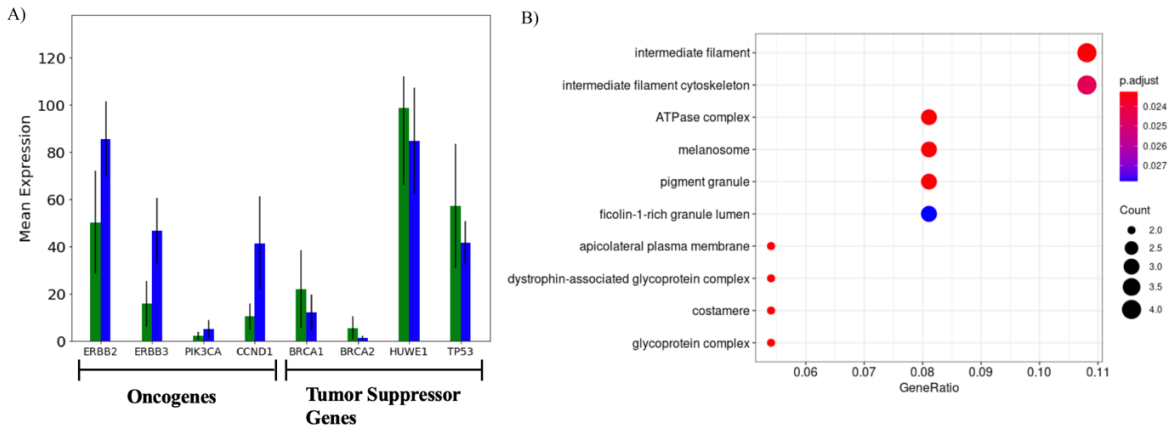

Figure 1.5: Differential expression across the single-cell phylogeny. A) Expression patterns of breast cancer specific oncogenes and tumor suppressor genes. B) Gene-set enrichment analysis of 42 genes that are differentially expressed across the phylogeny.

1.4 Discussion

Tumorigenesis is the gradual accumulation of somatic alterations in the genome, such as point mutations and copy number variations. From an evolutionary biology perspective, these alterations increase cellular fitness by providing proliferative advantage over cells without the alterations. Reconstructing the evolutionary history of a tumor critical for understanding ITH and developing personalized therapeutics. The gradual accumulation of alterations leads to change in molecular and organismal phenotype, and here we have provided a framework for detecting changes in transcriptome during tumorigenesis. Previous work has been done to reconstruct tumor phylogeny from single-cell DNA-sequencing data. In a parallel fashion, other studies have performed analysis of transcriptomic heterogeneity from tumor scRNA-seq data. This work builds on top of these methods by combining genomic and transcriptomic data to study heterogeneity during tumorigenesis, without the use of costly single-cell multi-omic sequencing data. However, there are some inherent limitations. The bisected experimental design can lead to sampling mismatch between the bulk DNA-sequencing data and the scRNA-seq data. Furthermore, somatic mutations

in rare sub-clones may be undetected due to using a bulk DNA-sequencing assay. Finally, somatic point mutations outside of coding regions are generally undetected due to the nature of RNA-seq, therefore the method relies on a small fraction of somatic mutations for inferring the phylogeny. Utilizing high-throughput 3'-based droplet scRNA-seq data can alleviate the aforementioned sampling mismatch by sampling more cells, however the per-cell coverage is 2 orders of magnitude less than that of plate-based scRNA-seq data which makes genotyping individual cells particularly challenging. For scRNA-seq datasets with lower coverage, building single-cell phylogenies using somatic copy number alterations is more likely and I leave this for future researchers to further investigate

We defined two sub-clones based on the reconstructed mutation tree and performed DE to define transcriptomic changes during tumorigenesis. Due to the low number of cells, the DE analysis was underpowered leading to 42 genes being detected as differentially expressed. Increasing the number of cells to the order of $\sim 10^2$ cells would alleviate the power issue. GSEA few enriched categories due to the limited number of genes detected, however we do recover GO terms associated with increased metabolism and increased cellular structural integrity owing to intermediate filaments.

CHAPTER 2: BENCHMARKING SINGLE-CELL AND SINGLE-NUCLEUS TRANSCRIPTOMICS[1]

2.1 Introduction

The identification and characterization of cell types from solid tissues and organs in the human body is the necessary basis for a comprehensive reference map of all human cells[31]. Single-cell RNA-sequencing (scRNA-seq) has emerged as a key tool to decompose complex tissues into cell types and states, and to investigate cellular heterogeneity. We have observed how scRNA-seq, paired with WES, can help us interpret somatic mutations and their putative downstream effects on tumorigenesis in Chapter 1. Profiling cellular heterogeneity using thousands of cells and creating tissue level cellular maps require efficient and scalable scRNA-seq protocols. The development of droplet-based approaches, such as Drop-seq, has enabled transcriptional profiling of thousands of cells in parallel[5], [32]. However, Drop-seq requires suspensions of intact single cells for library preparation which cannot be obtained for many tissues and cell types because of extra-cellular matrix that may be hard to digest, fragile cell membranes, unusual cell morphology, or large cell-size. This challenge may be addressed by adapting Drop-seq to single nuclei RNA-seq (DroNc-seq[12]). DroNc-seq obtains gene expression profiles from isolated nuclei which are more amenable for direct dissociation from tissues while maintaining membrane integrity. Both approaches can be used to characterize cellular composition of complex tissues. Comparisons of low-throughput, high-coverage single cell and single nucleus approaches suggest that both

1. Much of this chapter contains material from the paper: Systematic Comparison of High-throughput Single-Cell and Single-Nucleus Transcriptomes during Cardiomyocyte Differentiation (doi.org/10.1038/s41598-020-58327-6), which is currently published in Scientific Reports.

methods capture the cellular composition of heterogeneous samples to a similar degree[33], [34]. However, direct comparisons of Drop-seq and DroNc-seq on matched samples have been limited to cell lines[12] and, more recently, samples from mouse kidneys[35]. To establish a firm understanding of the differences and similarities of Drop-seq and DroNc-seq, it is necessary to compare these technologies across a spectrum of different biological conditions.

A crucial aspect of single cell RNA-seq approaches is to capture cellular heterogeneity associated with expression changes during dynamic processes, for example during differentiation. In this Chapter, we performed a systematic comparison of Drop-seq and DroNc-seq using time-course data from human iPSCs differentiating into cardiomyocytes (CMs). This allowed us to compare Drop-seq and DroNc-seq with respect to read depth, transcriptome composition, cell types detected, and cellular differentiation trajectories. These assessments are important for integrative analyses and interpretation of data produced using high-throughput single-cell and single-nucleus RNA-seq in general, and with Drop-seq and DroNc-seq in particular. In addition, we confirmed that inclusion of reads from intronic regions increases the sensitivity of DroNc-seq and improves resolution in identifying cell types. Next, we applied DroNc-seq to frozen *postmortem* human heart tissue to sample constituent cell types and compare them to CMs grown *in vitro* from human iPSC. This work was conceived as part of benchmarking experiments to establish the applicability of recent high-throughput single-nucleus RNA-seq for the Human Cell Atlas (HCA)[31]. By identifying differences and similarities between Drop-seq and DroNc-seq, this study will aid efforts such as the HCA that require the integration of single-cell and single-nucleus RNA-seq data from various tissues and laboratories into a common platform.

2.2 Methods

2.2.1 Bioinformatics of Drop-seq & DroNc-seq on iPSC-derived cardiomyocytes

To quantitatively assess the similarities and differences in transcription profiles from single-cell and single-nucleus RNA-seq, we performed Drop-seq and DroNc-seq, respectively, on cells undergoing iPSC to CM differentiation, following an established protocol[36]. To compare Drop-seq and DroNc-seq across samples with different cellular characteristics and degrees of heterogeneity, we collected cells from multiple time-points throughout the differentiation process (days 0, 1, 3, 7, and 15) (Figure 2.1A). For each technique, we obtained samples from two cell lines per time-point, except for time-point day 15 which contains cells from a single cell line. DroNc-seq also contains a single cell line for day 1. Drop-seq and DroNc-seq samples for each differentiation time-point were sequenced in a single run, with 150-200 million reads allocated per sample.

A total of 17 sequencing runs were performed over the course of the differentiation. Each sequencing run produced paired-end reads, with one pair representing the 12 bp cell barcode and 8 bp unique molecular identifier (UMI), and the second pair representing a 60 bp mRNA fragment. We developed a lightweight and efficient Snakemake[37] protocol (Figure 2.2) that takes a FASTQ file with such paired-end reads as input and produces an expression matrix corresponding to the UMI of each gene in each cell. The protocol initially performs FastQC[38] to obtain a report of read quality. Next, it creates a whitelist of cell barcodes using umi_tools[39] 0.5.3, which is a list of cell barcodes with at least 30k reads. Next, each paired-end read is combined into a single read where the read name contains the cell barcode and UMI extracted from paired end read 1, and the sequence content corresponds to paired end read 2. This is done for every paired end read and

placed into a single "tagged" FASTQ file. The tagged FASTQ file contains only the cell barcodes found in the whitelist. Finally, the protocol trims the ends of reads to remove polyA sequences and adaptors using cutadapt[40] 1.15. The tagged and trimmed FASTQ file is aligned to the human reference genome (version GRCh38) using the STAR[28] aligner version 2.5.3, which returns a BAM file sorted by coordinate. Next, we use featureCounts[41] version 1.6.0 to assign each aligned read to a feature on the genome.

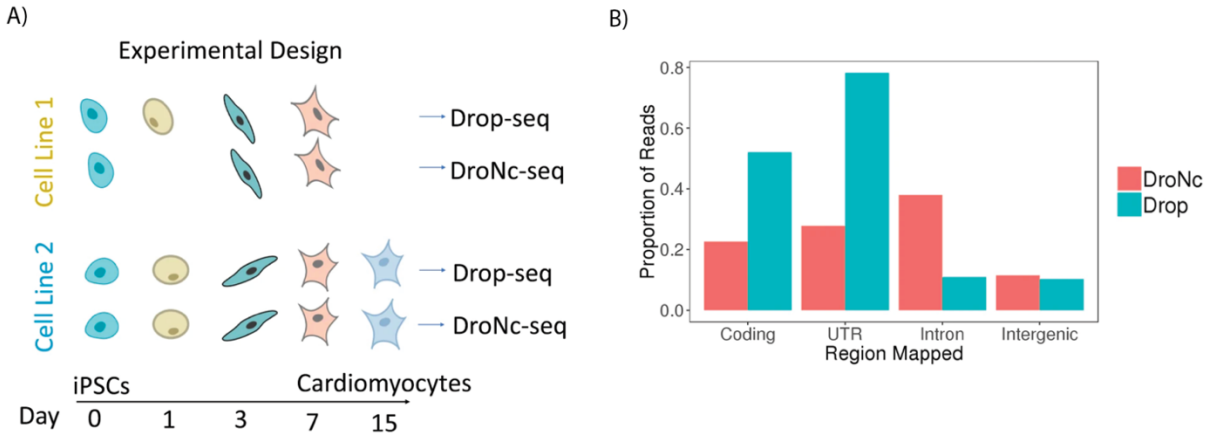

Figure 2.1: A) Bisected experimental design. Two technical replicates from two iPSC cell-lines were differentiated in-vitro and sampled at five timepoints. Drop-seq and DroNc-seq were ran in parallel on technical and biological replicates. B) Distribution of reads mapped across the genome from all RNA-seq experiments.

These features can be user-defined, but typically we are interested in quantifying the activity or expression of a gene. Typical RNA-seq bioinformatics pipelines will quantify a gene's counts by aggregating all reads that overlap with the gene's exons. Given the large proportion of DroNc-seq reads that map to the intronic region of genes (Figure 2.1B), likely due to nascent RNAs, we reasoned that incorporating these reads into quantifying gene expression can improve DroNc-seq data quality. In order to incorporate introns into the counting process, the UMI count of a gene was calculated as the sum of its exon and intron UMIs. GENCODE version 28 annotations contain exon features and gene features but do not contain intron features. To derive an intron annotation

file, we used exon and gene features. Exon regions were subtracted from gene regions (on the same strand) and the remainder was counted as the intron region for said gene. Then the expression level of a gene is given by the sum of the number of intron and exons.

In the final step of the pipeline, we use the count function from umi_tools to create a count matrix representing the frequency of each feature (gene) in the BAM file. The pipeline is available github.com/aselewa/dropseq_pipeline. A total of 17 count matrices were produced by this pipeline, 9 of which correspond to Drop-seq and 8 correspond to DroNc-seq.

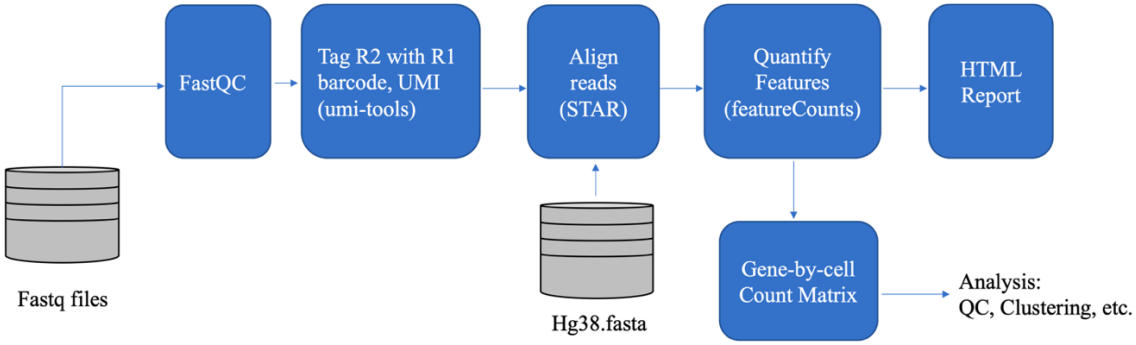

Figure 2.2: dropseqRunner pipeline schematic. FastQC is used to assess quality of paired-end reads, and then the read 2 is tagged with barcode and UMI information from read 1. STAR is used to align the reads and a count matrix is generated using featureCounts.

From each sequencing run, approximately 5000 cells were obtained with an average read depth of 30k – 40k per cell. Low quality cells were filtered based on the number of genes detected. A gene was considered detected in a cell if there was at least 1 UMI present. Cells with less than 400 genes and nuclei with less than 300 genes detected were removed. Low quality genes were also filtered if they were not detected in at least 10 cells, in order to reduce noise and computation cost. The total numbers of cells remaining were approximately 23,554 and 24,318 for Drop-seq and DroNc-seq, respectively. After filtering, all expression matrices from Drop-seq experiments were merged into a single expression matrix. The merging was done by taking the union of all genes. If a

particular dataset did not contain a gene that is expressed in another dataset, we set the expression level to zero in the first dataset. Similarly, all expression matrices corresponding to DroNc-seq were merged into a single expression matrix. Both merged matrices were processed and analyzed separately downstream. Seurat[42] was used to perform normalization, clustering, and cell type analysis. R scripts used for the analyses in this paper are documented at github.com/aselewa/czi.

2.2.2 Statistical analyses of Drop-seq and DroNc-seq

Normalization and Data Transformation

Having processed the single-cell count matrices and performed quality control, the first step in the clustering analysis of single-cell data is normalization and data transformation. The purpose of normalization is to remove technical variation across single cells, such as varying sequence depth or library size. A common strategy is computing counts-per-million (CPM), where the counts of each gene are divided by the cell's library size, and then scaled by 10^6. Another common method that builds on CPM is the transcripts-per-millon (TPM), where a gene's counts is divided by the gene's length and the cell's library size. Indeed, TPM is widely used with bulk RNA-seq assays where the full transcript length is sampled and therefore gene length will bias the gene counts. For droplet-based RNA-seq such as Drop and DroNc-seq, only the 3' end of the transcript is sampled and total gene counts are not biased by gene length. Therefore, we normalize the Drop-seq and DroNc-seq by only considering the library size, however, we use a scaling value of 10^4 which yields TP10k (transcripts per 10k) values. A pseudo-count of 1 was added to all scaled values followed by a data transformation. Traditional clustering algorithms, such as hierarchical clustering and k-means, cannot be directly applied to RNA-seq count data because they tend to follow an over-dispersed Poisson or negative binomial distribution. Therefore, we need to

transform the data in order to have a distribution closer to the normal distribution and one such simple transformation is the logarithm. After the log-transformation, the values were standardized, i.e. mean-centered and scaled such that each gene has unit variance. These log-normalized, and standardized data were used in downstream analyses to perform dimensionality reduction and reconstruction of differentiation trajectories.

Clustering methods that rely on Euclidean distance such as hierarchical clustering tend to perform poorly with single-cell RNA-seq methods due to the high sparsity levels imposed by the curse of dimensionality in single-cell datasets. One way to circumvent this is to reduce dimensionality and aggregate expression across genes whose expression is highly correlated. A popular technique to reduce dimensionality of RNA-seq data is principal component analysis (PCA). The principal components (PCs) describing most of the variation in the data are chosen as features for clustering analysis. Prior to PCA, Seurat calculates the gene dispersion vs. mean expression in order to obtain a subset of highly variable genes, which reduces the computational time of PCA compared with using the entire subset of genes identified in the experiment. Highly variable genes were selected based on a threshold of 1.5 for the dispersion level and a minimum mean expression level of 0.15 (on log scale) yielding 400 genes and 350 genes with Drop-seq and DroNc-seq, respectively. These highly variable genes were used to calculate PCs for Drop-seq and DroNc-seq data. The top 7 principal components, which explained 60% and 70% of variation for Drop-seq and DroNc-seq, respectively, were used to perform clustering.

Clustering Analysis

As mentioned above, hierarchical clustering is widely used for RNA-seq clustering, however the computational complexity does not scale well with the number of samples. In the case of hierarchical clustering, the computational complexity is $O(kn^2)$ where k is the number of desired

clusters, and n is the number of samples. High throughput scRNA-seq datasets such as those from Drop-seq and DroNc-seq are capturing cells from an unprecedented number of cells or samples. Other methods such as k-means has linear computational complexity $O(kn)$ and generally scales well, however, k-means relies on a fixed number of clusters and assumes equal shape and density of clusters. As an alternative, there has been much interest in applying graph-based clustering methods to scRNA-seq data. Graph-based methods have a smaller memory footprint, do not make assumptions about the number of clusters, and have been reported to outperform other clustering methods in many situations[43]. In graph-based methods, a k-nearest-neighbors graph is constructed were each node is a cell and the edges correspond to weights of similarity between two nodes. A greedy optimization algorithm called the Louvain method[44] is used for identifying communities or clusters in large networks. The algorithm works by portioning cells into a variable number of groups such that the Louvain modularity is maximized. The Louvain modularity is maximized by minimizing with-in cluster distance and maximizing between-cluster distance. The main parameters required by the Louvain method are k (for the KNN graph, not to be confused with k-means), and the resolution parameter which controls the granularity of the clusters. This method is implemented in the FindClusters method within Seurat and we apply this function to the top PCs for Drop-seq and DroNc-seq.

In order to adopt a criterion for choosing the resolution, we turned to the SC3 stability index[45]. The stability index compares a cluster at a particular resolution with all other clusters at all other resolutions. If the cluster being evaluated undergoes splitting for all changes of resolution, then it will be evaluated as relatively unstable. The stability index ranges from 0 to 1 denoting less stable to more stable, respectively. We ran the Seurat FindClusters algorithm for a range of resolutions from 0.01 (2 clusters) to 0.4 (12 clusters), and computed the stability index for each cluster and

each resolution (Figure 2.3). At each resolution, we took the mean stability across all clusters. The resulting stability profiles are shown below for DroNc-seq and Drop-seq. We used kernel regression in R (ksmooth function) to fit a line to the data. From these results, we see that mean stability is maximum near resolutions of 0.11 and 0.14 (red dashed line) for DroNc-seq and Drop-seq leading to 5 and 6 clusters, respectively. Importantly, the small differences in resolution parameters do not lead to any changes in number of clusters discovered and cell assignments to clusters. Therefore, we used a resolution parameter of 0.13 and *k* value of 30 to obtain 6 clusters in Drop-seq and 5 clusters in DroNc-seq. The clustering results were visualized with the Uniform Manifold Approximation and Projection (UMAP[46]), which produced a 2-dimensional visualization of the data. We also performed tSNE on the same data using a perplexity of 50 and found that UMAP captures more of the global structure in the data, as previously reported[47]. A minimum distance of 0.5 and 0.6 were used in UMAP for Drop-seq and DroNc-seq, respectively.

Differential Expression

In order to determine interpret the clusters and assign cell types or cell states, we performed differential expression analysis using the FindAllMarkers function and negbinom test in Seurat. This identifies differentially expressed genes between every two groups of cells using a likelihood ratio test of negative binomial generalized linear models. The Seurat's negbinom test yields relatively low false positive rates for differential expression analyses, compared with other parametric methods[48]. The p-values were adjusted for multiple testing using the Bonferroni correction. Furthermore, as we were only interested in upregulated genes as these will define the cell type, we ordered genes in each cluster, by their average log-fold-change (logFC) in descending order. Marker genes were identified based on functional annotations as these genes associated with cell types have a large fold-change in expression.

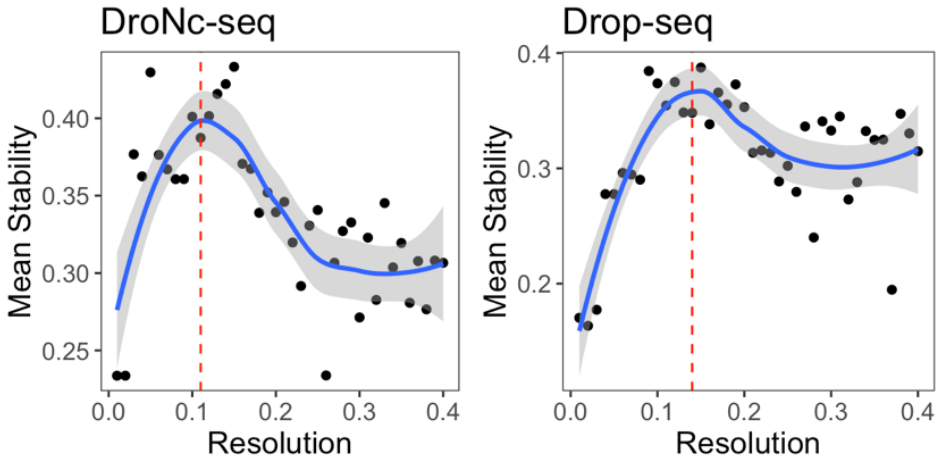

Figure 2.3: Cluster stability analysis for DroNc-seq and Drop-seq.

2.3 Results

2.3.1 Comparison of RNA-types between Drop-seq and DroNc-seq

Overall, Drop-seq shows a higher number of genes and transcripts detected compared with DroNc-seq, reflecting the greater abundance of transcripts in the intact cell, compared with the nucleus alone. For our analyses, we selected cells and nuclei with at least 400 and 300 detected genes (at least 1 UMI), respectively. After filtering, the mean number of genes detected per cell and per nucleus are 962 and 553, and the mean numbers of UMI per cell or nucleus are 1474 and 721 for Drop-seq and DroNc-seq, respectively. Based on the above cut-offs, we detected a total of 25,475 cells and 17,229 nuclei across all cell lines and time-points for Drop-seq and DroNc-seq, respectively. Both cell lines were present at each time-point in the filtered datasets. Using raw RNA-seq reads, we found that top expressed genes in Drop-seq comprised of mitochondrial and ribosomal genes, while the top gene in DroNc-seq was the non-coding RNA, MALAT1. We also compared genes detected in both protocols and found 273 genes that were only detected in DroNc-

seq. Out of these 273 genes 107 (39%) were long non-coding RNAs, which confirms that DroNc-seq is specifically sensitive to transcripts which often show strong nuclear localization.

To identify systematic differences in gene-specific detection rates between Drop-seq and DroNc-seq, we obtained differentially expressed genes (DEGs) between the two techniques for matched time-points and cell lines. As a comparison, we also performed differential gene expression analyses between time-points and between cell lines within each technique. We detected substantially more genes with differential expression between the two techniques than we observed between different time-points or cell lines (Figure 2.4A). The differentially detected genes directly reflect the sampling differences in cellular components for the two techniques. GO analysis on DEGs between Drop-seq and DroNc-seq revealed functional annotations associated with the sampling of different cellular components of the two techniques (Figure 2.4B). In particular, 5% of genes detected at higher levels in DroNc-seq were lncRNAs (compared to 1% in Drop-seq), while 20% and 6% of genes detected at higher levels in Drop-seq were mitochondrial and ribosomal transcripts, respectively (Figure 2.4C).

2.3.2 Incorporation of intronic reads

In addition to the differences in the number of genes detected in Drop-seq and DroNc-seq, DroNc-seq captures a significantly higher fraction of intronic reads compared with Drop-seq (Figure 2.1B). Up to 50% of the reads from DroNc-seq mapped to intronic regions, while for Drop-seq, only 7% of reads were intronic. This discrepancy between the two techniques is expected and likely caused by the sampling of unprocessed transcripts that are enriched in the nucleus. Intronic reads will be detected if the transcript was not fully processed before capture by the polydT primer. In addition, internal priming[49] on polyA stretches might lead to further sampling of introns. In order to

understand the sources of intronic reads in our dataset, we scanned the genome for polyA stretches that are at least 5 bp long, and counted their frequency within and around each read with 20 bp flanking regions. We found that approximately 40% of the intronic reads and their 20-bp flanking regions contained at least one polyA stretches and that these polyA stretches were specifically enriched towards the 3' end of reads (Figure 2.5A). This suggests internal priming as a contributing mechanism for intronic read sampling. RNA-seq reads aligning to introns have been used to quantify gene expression levels previously[34], [35],[50]. Indeed, incorporating intronic reads to quantify gene expression level improves the gene detection rate in DroNc-seq by ~1.5 times on average (Figure 2.5B). This increase in detection rate leads to recovery of gene expression for cells which would otherwise not be detected, as demonstrated by examples from mesoderm and cardiac genes (Figure 2.5C). These data suggest that inclusion of introns can be used to compensate for the smaller amount of nuclear RNA compared with whole cells. Accordingly, we incorporated intronic reads into our analysis pipeline to improve gene detection rates in DroNc-seq. After intron inclusion, we recovered 1.5 times more nuclei, bringing our total to 25,429 nuclei using a minimum of 300 genes detected per nucleus. In addition, the mean number of UMI per cell increased from 721 to 918, while the mean number of genes detected per cell increased from 553 to 672.

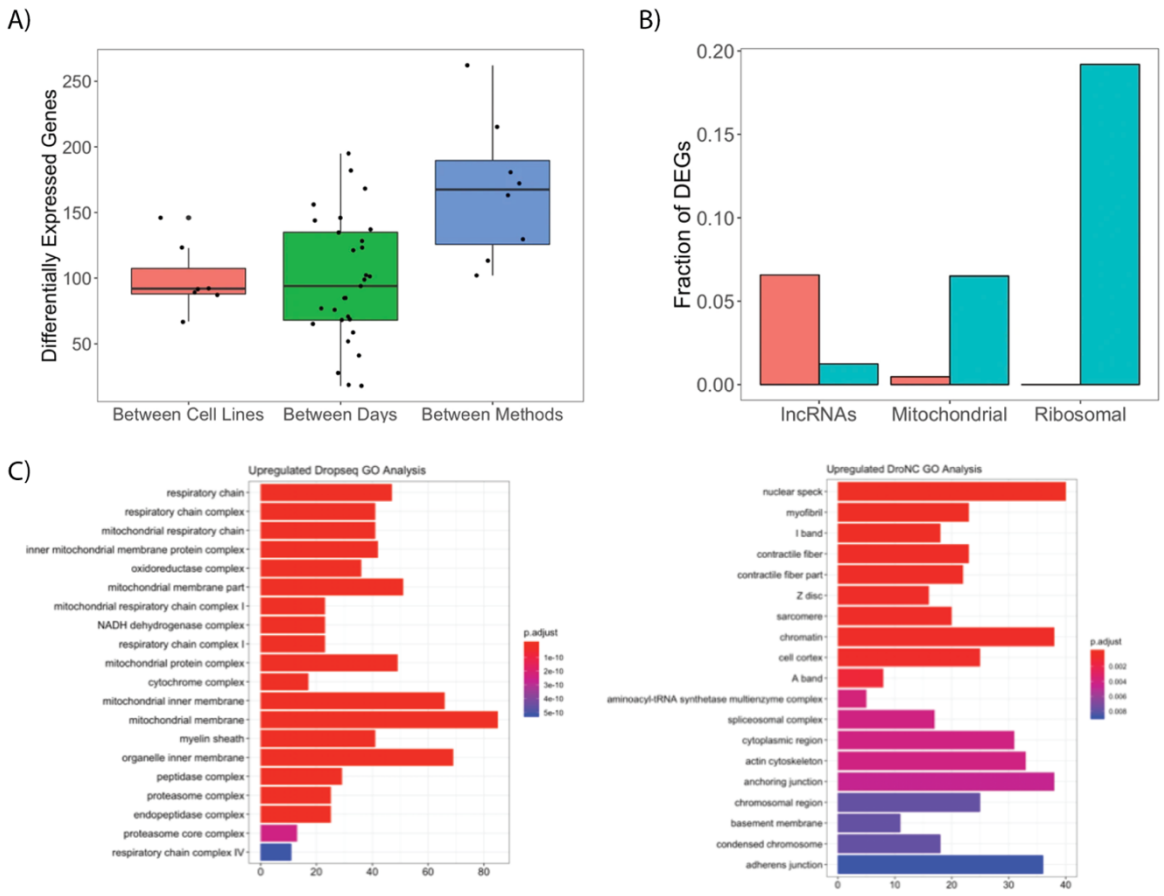

Figure 2.4: Systematic comparison of Drop-seq and DroNc-seq data. A) Differential expression between cell lines, days, and between the two modalities. B) Fraction of differentially expressed genes (DEGs) as long non-coding RNAs (lncRNAs), mitochondrial RNAs, ribosomal RNAs. C) Gene-set enrichment analysis on the DEGs between Drop-seq and DroNc-seq.

2.3.3 Communities detected in Drop-seq and DroNc-seq

Next, we tested if the differences between Drop-seq and DroNc-seq in the number of detected UMI and enriched gene sets lead to inconsistent detection of cell types and variation in the inferred differentiation trajectory. To infer cell types found with Drop-seq and DroNc-seq data, we performed clustering of cells separately for each technique as outlined in Section 2.2.1. Cell types were assigned to clusters based on comparison of genes that are significantly upregulated in the cluster to known marker genes. All genes were tested for differential expression using a negative

binomial likelihood ratio test within the Seurat package and p-values were adjusted for multiple testing using Bonferroni correction. For each cluster, we ordered genes by their average log-fold-change (logFC) in descending order to identify marker genes, as genes associated with cell type have a large fold-change in expression. Note that p-values (raw and adjusted) for all marker genes are small (adjusted $p < 10^{-5}$). We used the top marker genes for each cluster to identify cell type specific genes (Table 2.1). The cluster formed by cells from early time-points day 0 and day 1 contained pluripotent stem cells (Figure 2.6 A,B , 'iPSC', orange cluster), in agreement with the expression of characteristic markers such as DPPA4 (Figure 2.6 C,D). Cells harvested on day 3 mostly formed a separate cluster ('Cardiac progenitors', green cluster) composed of cells expressing markers concordant with cardiac progenitors (e.g. expression of EOMES (logFC=1.08), a mesendoderm progenitor marker gene). For days 7 and 15 the clusters of cells profiled by Drop-seq and DroNc-seq showed slight differences and we detected four clusters in Drop-seq compared to three for DroNc-seq, indicating that Drop-seq might be more sensitive towards detection. Drop-seq and DroNc-seq identified three clusters of ostensibly similar cell types. Two of these clusters contained cells predominantly expressing markers of CMs, including MYH6, TNNT2, MYL, and MYBPC3 (Figure 2.6 A,B, cyan cluster, 'Cardiomyocyte 1' and blue cluster, 'Cardiomyocyte 2'). We also detected a cell cluster that expressed cardiac markers alongside markers of other lineages (e.g. FOXA2 and TTR, pink cluster, 'Alternative lineage 1'). Figure 2.6 C and D show the expression of all markers discussed above. Drop-seq revealed an additional smaller cluster (purple, 'Alternative lineage 2', expression of FLT1) for which we did not find an equivalent cell population in DroNc-seq. These 'Alternative lineage' clusters might represent cells at intermediate stages, failures of differentiation, or differentiation towards alternative lineages. This heterogeneity and the detection of mesendodermal and endodermal cell populations, including

endothelial cells, is in agreement with previous scRNA-seq data obtained during iPSC to cardiomyocyte differentiation[51]. We also compared the pseudo-bulk profiles of iPSCs and CMs with external bulk assay RNA-seq and found Drop-seq to better capture bulk gene expression as expected (Figure 2.6E).

Markers	**Cell type**	**Prevalence (Drop-seq)**	**Prevalence (DroNc-seq)**	**Drop-seq Only Genes (top 5)**	**DroNc-seq Only Genes (top 5)**
DPPA4	iPSC	48.9%	52%	SFRP2, AC025465.1, ESRG, CACNAD2D3, BDNF-AS	RIMS2, RPL8, GOLGA4, EIF4A2, SET
EOMES APLNR	Cardiac Progenitor	23.3%	18.2%	CER1, LHX1, CYP26A1, IRX3, MT-TS2	GRIB2B, AL3365295.1, IL1RAPL2, KCNQ5, NRX3
MYH6 TNNT2	Cardiomyocyte 1	16.1%	5.6%	MYL3, NPPA-AS1, NPPA, ACTN2, TNNC1	AC012574.1, AC105233.5, MYO1D, ARHGAP42, CDK14
MYH6 TNNT2 AFP SERPINA1	Cardiomyocyte 2	4.2%	12.7%	AMBP, APOA2, SERPINF2, ITIH2, SERPINA5	KCNH7, ERBB4, ZBTB20, NRG3, KCNQ5
TTR FOXA2	Alternative Lineage 1	5.9%	11.3%	GATA3, S100A14, HHEX, FLIRT3, EPSTIL1	EWSR1, PTBP2, ZMYM2, LUC7L, LINC01876
CD34 SCARF1 FLT1	Alternative Lineage 2	1.4%	0%	CRHBP, GNG11, SOST, TFPI2, AC007319.1	None

Table 2.1: Marker genes used to assign cell type identity to each cluster.

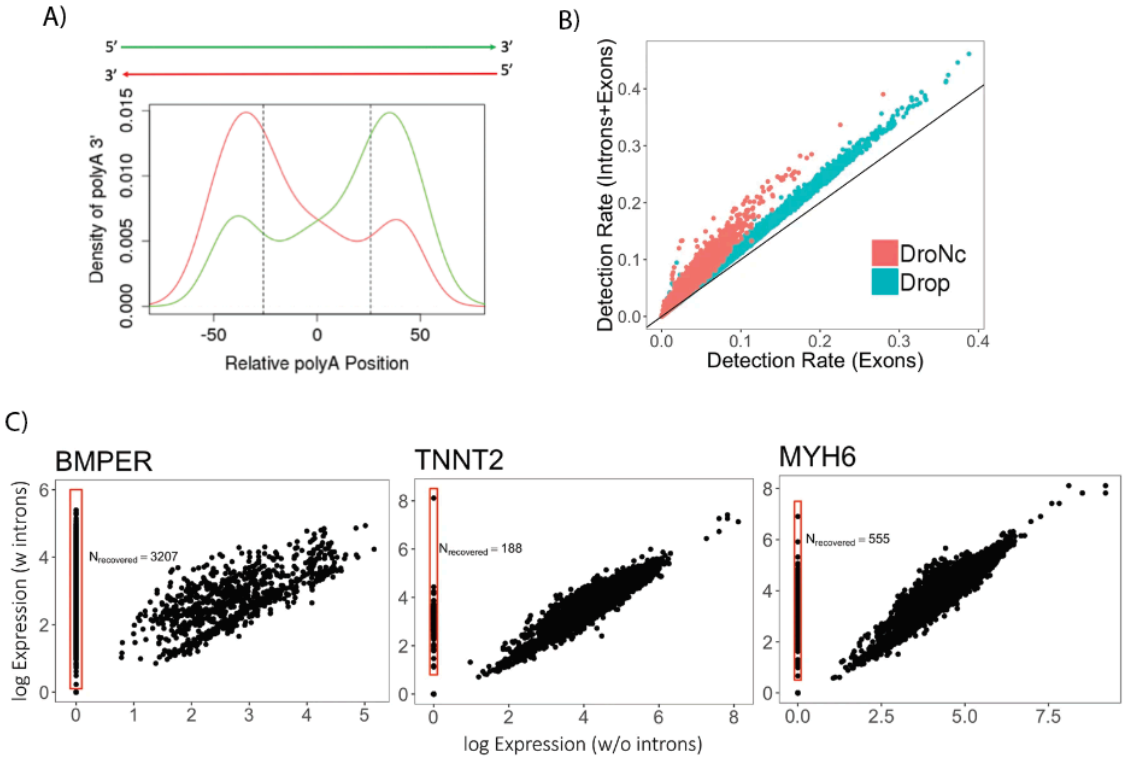

Figure 2.5: Effect of inclusion of intronic reads on gene expression. A) Strand-specific enrichment of polyA motifs for reads mapped. B) Gene detection rate with and without intronic reads for Drop-seq and DroNc-seq.

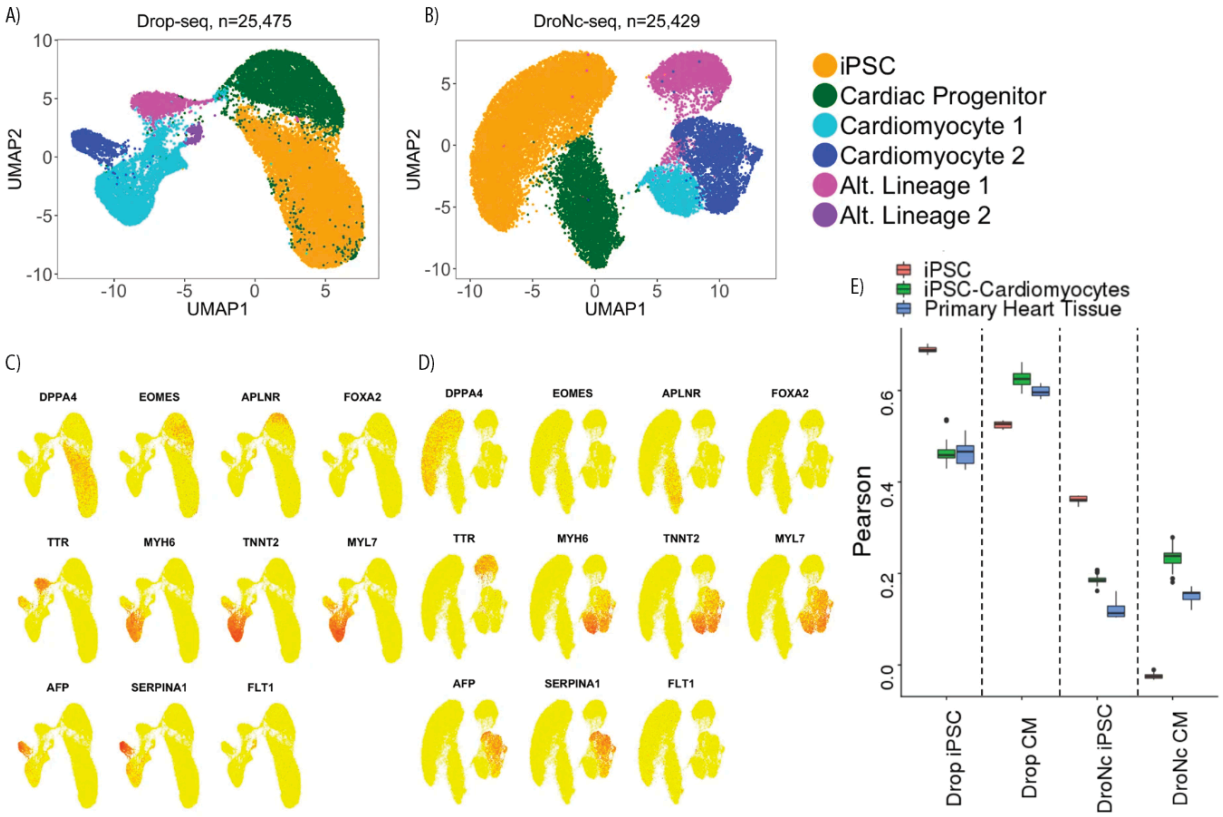

Figure 2.6: Results of clustering analysis on Drop-seq and DroNc-seq. A,B) UMAPs with color representing cell types. C,D) UMAP with marker gene expression overlaid. E) Pearson correlation of iPSC and CM pseudo-bulk with bulk assay data.

2.4 Discussion

Building a cell atlas of the human body requires the expression profiling of all human tissues from a range of different samples, including tissues that are hard to dissociate, composed of fragile cells, and frozen specimens, all of which are incompatible with single-cell RNA sequencing. As an alternative, DroNc-seq, a high-throughput single-nucleus RNA sequencing protocol, has the potential to reveal tissue heterogeneity, at scale, based on *nuclear* RNA, and is being increasingly used to profile primary tissue at high throughput. However, it is unclear how DroNc-seq compares with earlier single-cell RNA-seq protocols like Drop-seq across a range of different cell types and tissues. Previous studies have performed cell type comparisons using nuclear vs. whole-cell RNA

using full-length mRNA sequencing assays at low throughput[33], [34]. Drop-seq and DroNc-seq have been compared using adult mouse kidneys cells[35]. We performed a direct comparison of high-throughput, single-cell (Drop-seq) and single-nucleus (DroNc-seq) RNA-seq using iPSCs differentiating into CMs. This study enabled us to compare cell type detection, transcriptome profiling and infer cellular communities with two complementary high-throughput techniques, using an *in vitro* model of CM differentiation.

As expected, the number of UMIs per nucleus in DroNc-seq are lower than those for cells in Drop-seq. Consequently, the gene detection rate in DroNc-seq was significantly lower than for Drop-seq. However, given the high number of reads in DroNc-seq that mapped to intronic regions we reasoned that inclusion of such reads might increase the gene detection rate. Indeed, intron inclusion significantly increased the sensitivity of DroNc-seq and generally improved cluster separation in agreement with previous studies[33]–[35],[50]. In particular, the separation between iPSCs and Cardiac Progenitors became clearer after the inclusion of introns, while the change in separation for other clusters was negligible. We also found that the inclusion of introns increased gene detection rate in single nuclei samples. Of note, a significant proportion of the intronic reads seems to originate not from transcripts primed at the 3' end but from direct priming to polyA stretches in introns[49]. While such reads still scale with the expression level of a transcript, the assumption that transcript levels are uniquely quantified by a single UMI may be violated in these cases.

Given the difference in input material, i.e., cellular vs. nuclear RNA, it is not surprising that we found a significant proportion of genes that are differentially expressed between Drop-seq and

DroNc-seq samples. Some of the most highly enriched sets of genes reflected the technical differences between the two technologies. Genes specifically enriched in Drop-seq are ribosomal and mitochondrial. DroNc-seq presumably loses these transcripts that are predominantly localized in the cytoplasm. Conversely, as a class, lncRNAs are enriched in DroNc-seq which agrees with the nuclear localization of many of them.

Expression profiles in Drop-seq and DroNc-seq confirmed the differentiation of iPSCs into CMs and revealed major cell types found within the *in vitro* differentiation model of iPSC-CMs. These data also confirmed heterogeneity observed during differentiation. Drop-seq and DroNc-seq detected a population of cardiac progenitors with cellular prevalence 23.3% and 18.2%, respectively. They also both detected two clusters representing CMs: cardiomyocyte 1 (16.1% and 5.6% prevalence) and cardiomyocyte 2 (4.2% and 12.7% prevalence). Both methods also revealed a population of cells, 'Alternative lineage 1', that might represent alternative fate or that failed to reprogram fully, which accounted for 5.9% and 11.3% of all cells in Drop-seq and DroNc-seq, respectively. The presence of non-CMs during late-stage is expected for the *in vitro* differentiation model and has been observed previously[51]. Accordingly, the proportion of cells differentiating into CMs expressing TNNT2, assessed by FACS, varies widely between 20-80%[36]. Based on our cell type assignment in Drop-seq data, we obtained 28% and 29% cardiomyocytes on day 7 for the two cell lines and 70% CMs on day 15 for cell line 2, which fall within the expected range.

Drop-seq revealed an additional smaller cluster (purple, 'Alternative lineage 2', expression of FLT1 and comprising 1.4% of the total population) for which we did not find an equivalent cell population in DroNc-seq. The reasons behind the failure of DroNc-seq to identify the small

fraction of cells identified as ‘Alternative lineage 2’ in Drop-seq may be due to the lower capture rate of DroNc-seq (mean number of detected genes was 672) compared to Drop-Seq (mean number of detected genes was 962) which might result failure of the clustering approach to resolve this sub-population in DroNc-seq, or due to the preferential loss of the particular cell type arising from DroNc-seq’s nuclei dissociation protocol. The mean number of genes detected in this subpopulation in Drop-seq was 1032, representing the cluster with the highest gene detection rate. It is possible that this facilitated the detection of this cluster in Drop-seq while the lower detection rate in DroNc-seq combined with the small number of cells corresponding to this cluster in the sample lead to the loss of this population during clustering. However, we cannot rule out specific loss or selection biases for of the cell type introduced during DroNc-seq sample preparation. This comparison of Drop-seq and DroNc-seq demonstrates the capability of DroNc-seq in dissecting the multicellular environment within a complex tissue such as the heart, which would otherwise not be possible with Drop-seq. We expect that DroNc-seq will be used to perform high-throughput transcriptomic profiling of tissues for which it is difficult to obtain suspensions of intact single cells and aid in initiatives such as the Human Cell Atlas and the Human Tumor Atlas.

CHAPTER 3: SINGLE-CELL EPIGENETICS FOR INTERPRETING COMPLEX TRAITS

3.1 Introduction

Common genetic variants, or SNPs, can affect molecular and organismal traits through disruption of regulatory mechanisms. Genome-wide association studies (GWAS) have identified thousands of loci for hundreds of diseases and physical traits, including cardiac diseases[3]. Cardiac diseases impose a significant health burden world-wide. GWAS have provided important insights into the biology of many complex cardiac traits by identifying putative disease genes and pointing to disease-relevant pathways. For instance, pathogenic variants associated with LDL cholesterol levels and coronary artery disease (CAD) have helped redesign therapeutic strategies for these diseases[52]. Recently, large-scale GWAS have been applied to study electrocardiographic traits such as PR, RR, and QT intervals, and complex arrhythmias such as atrial fibrillation (AF). However, the task of translating GWAS hits into molecular mechanisms that explain disease risk and etiology still poses several formidable challenges for the field.

Identifying the causal variants that are underlying the associations found in GWAS is a common problem and is of great interest for dissecting the genetic architecture of complex traits. Because of extensive linkage disequilibrium (LD) in the human genome, a single causal variant can generate spurious associations in nearby SNPs that are in LD. Furthermore, the index or lead SNP, defined as the SNP with most significant association, is not always the causal SNP due to differential power to detect associations across SNPs as a result of variance in allele frequencies. One common way to address this challenge is to identify functional variants in disease-associated loci, which is made possible by functional genomics data that annotate the non-coding genome

which houses the majority of strong GWAS variants. It is believed that most of these non-coding variants are regulatory variants[53] that alter the activity of a gene-product by modulating the abundance of the gene-product through regulatory mechanisms. Given epigenomic annotations from functional genomics data, candidate variants in trait-associated loci can be prioritized based on their locations within regulatory elements. Bayesian fine-mapping approaches estimate the probability that a variant is causal and utilize SNP prior probabilities based on such functional annotations. Despite these advances, narrowing down to a single or few causal variants remains challenging. Regulatory elements and their activities are highly dynamic across tissue, cell-type/state, and across environmental stimuli[4]. Large consortia such as ENCODE provide epigenetic data spanning multiple primary tissues which helps bridge the gap, however cell-type resolved epigenetic data is still lacking for most tissue types.

We hypothesized that identification and characterization of cell-type-specific regulatory elements active in the human heart *in vivo* facilitates the interpretation of complex cardiac traits. Single-cell genomics approaches are powerful techniques for such studies because they can deconvolve complex cellular mixtures into their constituent cell types and characterize them at molecular level. Large-scale mapping of cell-types and -states revealed a remarkable cellular diversity in the adult human heart, demonstrating that normal cardiac function depends on the finely tuned interplay of multiple cell-types. This complexity highlights the importance to identify the cell types in which genetic variants exert their effects.

To test our hypothesis and characterize the distribution of GWAS SNPs within cell-type-specific regulatory features, we generated a cell-type resolved atlas of chromatin accessibility and gene expression in the human heart. We used these data to define cell-type-specific regulatory features. We observed strong enrichment of heart-related GWAS variants in cell-type-specific regulatory

elements. We used cell-type specific chromatin accessibility as functional priors in statistical fine mapping of 122 blocks in atrial fibrillation GWAS summary statistics.

Statistical fine-mapping approaches such as our own are focused on narrowing down on the underlying causal variants. However, narrowing down to a single or few causal variants remains challenging. For example, in a high-powered GWAS of type 2 diabetes (T2D), only in a quarter of associated loci, the credible sets (the union of all SNPs that, with high probability, contains the causal signal), have less than 10 SNPs[54]. One way to address this challenge is to combine signals from SNPs likely targeting the same genes. The intuition is that, after fine-mapping a locus, we may still have a large uncertainty of exact causal variant, but if most of candidate variants target the same gene, then that gene is likely to be the risk gene at the locus. This strategy will both increase the statistical signal and provide more interpretable results. We thus developed a computational procedure that summarizes statistical evidence from GWAS fine-mapping at the level of genes. Applying this approach to AF GWAS, we identified multiple genes that are likely altered by AF-associated variants, such as TBX5 and ETV1, the transcription factors that are believed to be key for cardiomyocyte development and function. Furthermore, the genes identified are enriched among the differentially expressed genes in cardiomyocytes.

3.2 Methods

3.2.1 Application of scATAC-seq and scRNA-seq on adult heart tissue

Tissue samples were obtained from the left and right ventricles (LV and RV), the interventricular septum, and the apex of three adult male donors. We isolated nuclei and purified these using fluorescence-activated cell sorting (FACS) to remove debris and minimize contamination from ambient RNA. We generated a total of 12 snRNA-seq and 12 scATAC-seq libraries using the

Chromium (10x Genomics) platform followed by high-throughput sequencing. Single-cell RNA-seq data were processed using dropseqRunner as described in Section 2.2.1. We utilized intronic and exonic reads when quantifying gene expression as discussed in Chapter 2. We combined all 12 expression matrices into a single Seurat object and kept track of the corresponding metadata such as donor and anatomical region. To account for low-quality nuclei, we removed barcodes that contained less than 1000 UMI. We also used DoubletFinder[55] v2.0.3 with hyper-parameters pN = 0.015 and pK = 0.005 to estimate doublets, which works by generating *in-silico* doublets and performs clustering to identify nuclei that fall in the neighborhood of the simulated generated doublets. After quality control, we retained a total of 49,359 nuclei with a range of 1,506-7,356 nuclei per sample.

We processed the snATAC-seq in an analogous fashion. FastQ files from 12 sequencing experiments were individually processed using CellRanger-atac v1.2.0. We used the command "cellranger-atac count" to align the FastQ files to human reference genome Hg38, followed by marking and removing duplicate reads, and produced a fragments file containing the mapped location of each unique fragment in each nuclei. We used ArchR[56] v0.9.5 to further pre-process the data and perform downstream analyses. ArchR (Analysis of Regulatory Chromatin in R) provides an efficient and scalable platform for complex snATAC-seq analyses such as dimensionality reduction, clustering, marker feature identification, transcription factor (TF) enrichment, TF foot printing and other utilities. ArchR takes as input BAM or fragment files, both of which are returned by CellRanger-atac, and stores the single-cell data in a compressed and memory efficient format on disk using HDF5. ArchR outperforms other packages such as snapATAC[57] and Signac[58] in speed and memory usage across routine analysis such as clustering and marker feature identification.

Using ArchR, we converted the fragments file into a tile matrix, which is a bin-by-barcode Tn5 insertion count matrix, using a bin-size of 500 bp. The tile matrix is preferable to a peak-by-cell matrix, which is commonly used in other tools as a starting point, because rare cell-types may have cell-type specific peaks that are not captured by bulk peak calls. ArchR also generates a gene score count matrix which aggregates the Tn5 insertion counts at each gene's promoter site for all genes in GENCODE v29. The gene score count matrix is also called the gene activity matrix, and it is analogous to gene expression. Indeed, gene expression and gene activity are highly correlated. To account for low quality nuclei, we kept nuclei with at least 5,000 unique fragments and a TSS enrichment score of 6. We also used ArchR's doublet removal scheme with default parameters, which is based on in-silico doublet generation, to remove nuclei with a doublet enrichment score greater than 1. After quality control, we retained a total of 26,714 nuclei with a range of 1,000-4,873 nuclei per sample.

3.2.2 Inferring cell-type resolved regulatory programs from snATAC-seq

In this section, we outline the procedure for inferring cell-types, using both snRNA-seq and snATAC-seq, and then define open chromatin regions (OCRs) for each cell-type. We performed normalization, dimensionality reduction, and unsupervised clustering on snRNA-seq and snATAC-seq data in order to map cell-types. For snRNA-seq, we used Seurat's workflow as described in Section 2.2.1. Briefly, Seurat begins with converting counts to log2 TP10k values using the NormalizeData function. Next, we found the top 2000 variable genes using FindVariableGenes and used these genes as input features for PCA. We computed the top 30 principal components (PCs) for each cell and used these for downstream analyses. We observed batch effects due to different donors, and we decided to correct this batch effect because we are

not interested in biological differences between donors in this study. To this end, we used the RunHarmony function from the Harmony[59] v1.0 package with default parameters in order to regress out the donor variable from the PCs. Next, we used the FindClusters with a resolution of 0.2 on the harmony-corrected PCs to define clusters. We computed the corresponding UMAP to visualize the harmony-corrected PCs in two dimensions and used the set of canonical markers listed in Table 3.1 in order to map clusters to cell types.

We also performed cell-type mapping for snATAC-seq using ArchR. We performed dimensionality reduction on the tile matrix using the top 20,000 bins in terms of count across all cells. We used the function addIterativeLSI with 2 iterations in order to perform latent semantic indexing (LSI) on the snATAC-seq tile matrix and retained the top 50 LSI vectors. Similar to snRNA-seq, we observed batch effects across different donors, and opted to remove this effect using the RunHarmony function. We used addClusters with resolution = 0.2 in order to cluster nuclei based on the harmony-corrected LSI vectors. addUMAP with min.dist = 0.4 was used to compute a 2-dimensional representation of the harmony-corrected LSI vectors. We visualized gene activity scores using the same marker genes (Table 3.1) as in snRNA-seq in order to assign clusters to cell-types.

Cell Type	Marker
Cardiomyocytes	TTN TNNT2 MYH7 MYBPC3
Pericyte	RGS5 ABCC9
Smooth Muscle	MYH11 TAGLN
Fibroblast	DCN PDGFRA
Endothelial	PECAM1 VWF
Neuronal	PLP1
Lymphoid	CDA8 LCK
Myeloid	CD14 FOLR2

Table 3.1: Marker genes used to identify cell-types in snRNA-seq and snATAC-seq.

Having defined the cell-types, we next outline the procedure for calling OCRs. Insertion read counts were aggregated across all cells in each cell-type to form a cell-type pseudo-bulk and peak calling was performed on each cell-type pseudo-bulk. Using the function addReproduciblePeakSet in ArchR in conjunction with MACS2, a union set of 352,900 OCRs were called in total across all cell-types at 10% FDR. This set of peaks is our basis set of regulatory features and was used for all downstream analyses.

In order to discover cell-type specific OCRs, a single-cell insertion count matrix was created using the function addPeakMatrix in ArchR. Cells were grouped into their respective cell-types and differential accessibility (DA) analysis was performed in a one-vs-all fashion. To perform DA, we used getMarkerFeatures in ArchR with default parameters, which uses the Wilcoxon rank-sum test

on the log-normalized insertion count matrix. In order to control for technical variation, cells from the cell-type group and the remaining group are matched in terms of TSS enrichment and number of fragments. Using an FDR < 1% and a $\log_2$ fold-change > 1, we found about 47%, or about half, of the union set to be cell-type specific. The features obtained from DA are not guaranteed to be cell-type-exclusive because DA only yields features that are more accessible in one cell-type relative to all the others combined. Therefore, it possible for a feature to be up-regulated in two cell-types. However, we find that the set of DA OCRs for each cell-type are approximately disjoint, with an upper limit of 6% sharing between the OCRs of any two cell-types. We also defined cell-type-shared peaks and peaks that were not in the cell-type of interest. To do this, we started with the union set of peaks, and removed peaks that were cell-type specific, according to the DA analysis above. The remaining peaks were considered shared or not relevant to this particular cell-type.

Having defined the cell-type-specific OCRs, we sought to characterize the transcription factors (TFs) that are potentially actively binding in each cell-type. To this end, we obtained sequence motif coordinates of approximately 870 human TFs from CisBP[60]. We used ArchR to estimate the enrichment of each TF within the specific OCRs of each cell-type using a simple hypergeometric test. We realized that this approach will lead to high enrichment of degenerate motifs by chance. To narrow down the list of TFs that are potentially active, we performed a correlation analysis between the motif accessibility of each TF, and the gene activity score of each TF. We expect that TFs being transcribed should also be binding to their motif sequence in the genome. The correlation was performed across meta-cells, which are pseudo-bulk aggregates of approximately 100 nearest neighbors. Using an enrichment with FDR < 1% in at least one cell-

type and |correlation| > 0.5, we find a total of 158 TFs that are putatively active. We reference these TFs later when investigating the functional effects of GWAS SNPs.

3.2.3 Fine-mapping GWAS summary statistics using functional priors

In this section, we outline the procedure of obtaining GWAS summary statistics for multiple traits and performing enrichment analysis of cell-type-resolved OCRs. Finally, we outline the fine-mapping procedure based on functional priors generated during the enrichment analysis. We obtained harmonized GWAS summary statistics from the IEU OpenGWAS project for cardiovascular and non-cardiovascular disease traits, as well as non-disease traits. Each set of summary statistics were further harmonized by utilizing the finemappeR R package developed and maintained by Alan Selewa and is available at github.com/aselewa. The provided script first cleans the summary statistics by removing SNPs with missing values, SNPs on non-autosomal chromosomes, and indels. Utilizing LD blocks generated by ldetect[61], we assign each SNP to one of 1703 approximately independent regions. Finally, we add population-matched genotype data to each SNP by utilizing the 1000 Genomes Project[62]. These out-of-sample genotypes will be used for estimating LD between SNPs for fine-mapping.

Next, we estimated the fold-of-enrichment of cell-type-specific OCRs in the summary statistics of multiple traits, including AF. For this, we employed a Bayesian hierarchical model TORUS[63]. TORUS takes as input GWAS summary statistics and genomic annotations (here we use OCRs), and for each annotation outputs enrichment estimates that correspond to estimates from a logistic regression: the additive change in log odds for a variant being causal, conditioned on all other annotations being held constant. Furthermore, TORUS also constructs SNP-level prior probability

of causality during the enrichment estimation, which we will utilize for functionally-informed fine-mapping.

Based on the enrichment patterns, we opted to perform fine-mapping on AF using functionally informed priors using the union set of OCRs. Fine mapping was performed using a summary statistics-based version of the "Sum of Single Effects" model (SuSiE[14]). In particular, we used the susie_rss function to fine-map each LD block, which takes GWAS z-scores and an LD matrix for the SNPs in the block. Because only summary statistics are available publicly, we used out-of-sample genotype information to construct LD matrices. In particular, we used the 1000 Genomes European panel of genotypes to compute LD. We ran SuSiE in L = 1 mode, which returns a single causal event for each LD block and is fairly robust to mismatching LD patterns. The prior probability of each SNP being causal was generated by TORUS using a joint-model of the following annotations: CM specific OCRs, CM shared OCRs, non-CM OCRs, UCSC conserved/coding. We fine-mapped a total of 121 LD blocks for atrial fibrillation, each containing at least 1 SNP at genome-wide significance ($P < 5 \times 10^{-8}$).

3.2.4 Gene-level summary of fine-mapping results

While statistical fine-mapping helps to identify putative causal variants, it does not directly implicate risk genes. This task is made difficult by several complications, including the possibility of distal regulation by enhancers, and the fact that a variant/ enhancer may affect multiple nearby genes. Existing work to assign target genes often focus on the top GWAS variant in a locus, and prioritizes the targets by distance, chromatin interactions informed by Hi-C or similar datasets, or by eQTLs. In reality, however, there is often considerable uncertainty of causal variants, as suggested by our own and other works. We developed a novel procedure to address these

limitations. It is directly informed by variant-level fine-mapping analysis, and aggregates information of all SNPs potentially targeting the same genes, weighted by their PIPs, considering multiple ways a SNP may affect its target gene. We used the posterior inclusion probabilities (PIPs) generated by SuSiE to calculate a gene-level PIP. Let Z_g be an indicator variable describing whether gene g is causal (Z_g=1) or not (Z_g=0) for the trait. Assuming a single causal SNP per locus, the gene PIP is then given by:

$$P(Z_g=1|D) = \sum_{SNP} P(Z_g=1|\gamma_{SNP}=1)\ PIP_{SNP}$$

where γ_{SNP} is the indicator variable for whether a SNP is causal or not, and *D* is the GWAS summary statistics. The term $P(Z_g=1|\gamma_{SNP}=1)$ can be interpreted as the probability of a SNP interacting with gene g . Finally, PIP_{SNP} is simply the PIP from fine-mapping. Since $\sum_{SNP} PIP_{SNP}=1$ for a single locus/block, the gene PIP has an upper-bound of 1. In cases where multiple loci/blocks are linked to a single gene, the gene PIP may exceed 1, in which case it can be interpreted as the expected number of signals linked to gene g.

The term $P(Z_g=1|\gamma_{SNP}=1)$ can be approximated by considering the location of the SNP with relation to the gene g, as well as functional genomics data that provide evidence for interaction a SNP with gene g. To do so, we assign weights $w_{SNP,g}$ to each SNP and gene, and then normalize by the total sum of the weights over all genes to estimate $P(Z_g=1|\gamma_{SNP}=1)$. If a SNP is in the exon, UTR, or promoter of a gene, we assign a full weight $w_{SNP,g}$ = 1. Furthermore, if a SNP is in an enhancer that is linked to a gene's promoter via promoter-capture HiC data and/or co-accessibility in the relevant tissue, then we also assign a full weight. If a SNP is not linked to any gene via the

methods above (e.g. in intronic and intergenic with no pcHiC or co-accessibility links), then we use a distance-based weighting to all genes with 1Mb:

$$w_{SNP,g} = e^{-d_{SNP,g}/10^5}$$

With all the weights estimated, probability of SNP-gene interaction is given by:

$$P(Z_g=1|\gamma_{SNP}=1) = \frac{w_{SNP,g}}{\sum_g w_{SNP,g}}$$

As a result, a SNP's PIP distributed to all genes linked in a weighted fashion determined by functional evidence and distance in cases where the former doesn't exist.

3.3 Results

3.3.1 snATAC-seq and snRNA-seq identify eight major cell-types

Unsupervised clustering on dimensionality-reduced snRNA-seq and snATAC-seq revealed eight distinct clusters. Utilizing the marker genes in Table 3.1, we were able to assign a cell-type to each cluster. The cell-types were exactly matched across the two modalities (Figure 3.1 A,B). Perhaps the central cell-type to the heart is the cardiomyocytes (CMs), the beating cells of the heart. The CM cluster showed high expression and gene activity of ventricular CM markers TNNT2, MYBPC3, and MYH7, as well as a depletion of atrial CM marker NPPA. This is expected given that we did not sample the atrium. In addition to CMs, we identified endothelial cells, fibroblasts, pericytes/smooth cells, a small population of neuronal/pace-maker cells, and an immune compartment. The top three cell-types in terms of proportion were CMs, endothelial cells, and fibroblasts (Figure 3.1, A,B side-panel), which constituted approximately 70% of cells. A discrepancy was observed in the ranking of cell-type proportion between RNA-seq and ATAC-seq, which is likely due to a combination of uneven sampling of different heart regions and the fact that different heart regions have different cell-type proportions. We further assessed the

concordance between snRNA-seq and snATAC-seq cell-type labels in two ways. First, genome-wide gene scores (scATAC-seq) and transcript levels (snRNA-seq) were highly correlated between matched clusters (Figure 3.1 C). Second, we performing label transfer as part of the Seurat v3 package where the cell-type assignment of each cell from snRNA-seq is transferred to its nearest-neighbor in the snATAC-seq data. We find strong concordance between the two modalities as the cell-type labels from snRNA-seq unambiguously transferred to the snATAC-seq data (Figure 3.1 D).

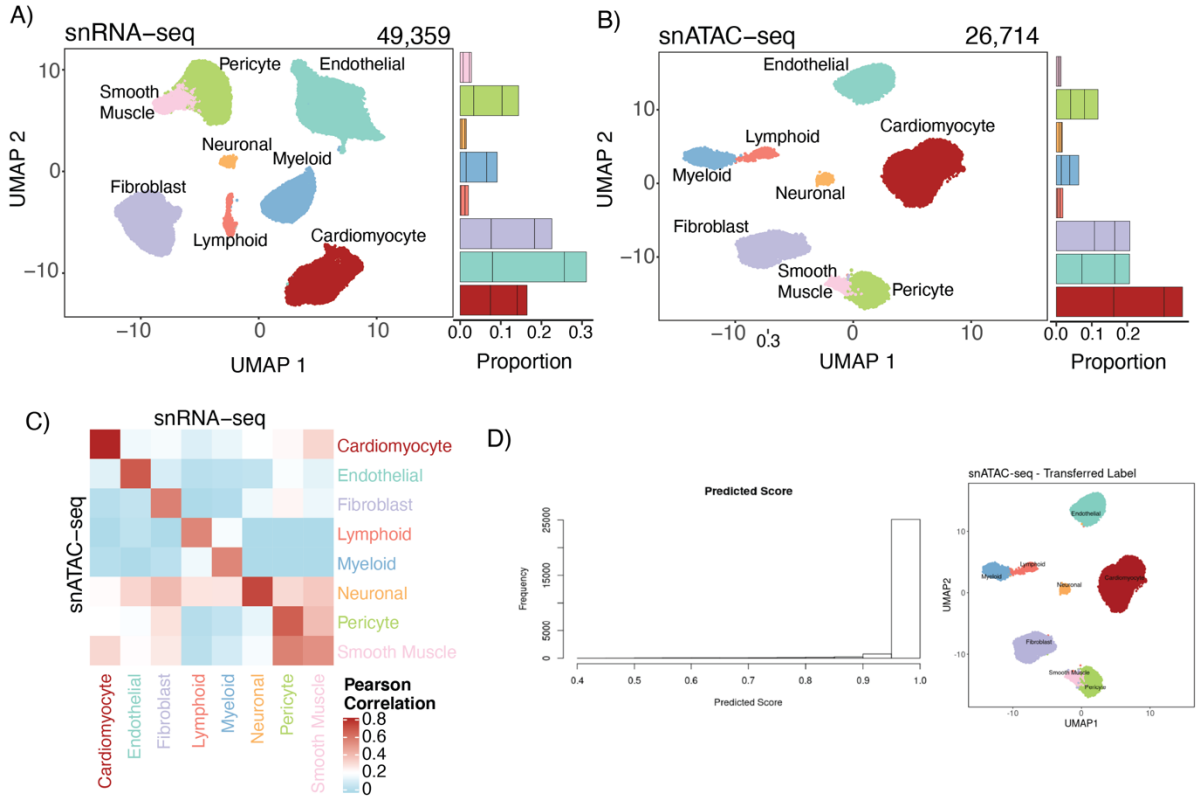

Figure 3.1: Cell-type mapping in snRNA-seq and snATAC-seq. A,B) UMAPs of snRNA-seq and snATAC-seq, respectively, with color indicating cell types. Barplots on the right of each UMAP show the proportion of each cell-type across the three donors. C) Pearson Correlation of clusters in snRNA-seq and snATAC-seq. D**)** Left: Histogram of label transfer scores. Right: UMAP of snATAC-seq with labels transferred from snRNA-seq.

3.3.1 Regulatory architecture of the human heart

Cell-type-specific chromatin accessibility is a powerful means to identify regulatory features associated with cell-type-specific gene expression programs. We used snATAC-seq data to identify cell-type-resolved OCRs and detected 44,997-150,000 OCRs per cell type using the peak caller MACS2. OCRs from all cell types were combined into a single union set comprising 352,904 OCRs. This procedure ensured that OCRs from rarer heart cell types (e.g., smooth muscle and myeloid cells) were represented in our analysis. Simple clustering of regions based on their accessibility profile revealed clear cell-type-associated accessibility patterns (Figure 3.2A). Based on this observation we identified OCRs with significant cell-type-specific accessibility for each of the eight cell types, as outlined in Section 3.2.1. In total, 173,782 (49%) OCRs showed cell-type-specific accessibility, we refer to the remaining 179,122 (51%) OCRs as shared. Cell-type-specific OCRs tended to include more promoter distal regions compared to shared OCRs (Figure 3.2 B). Visualizing the distribution of distance to the nearest TSS, we find that the shared category has a small peak at ~1kb, indicating the presence of promoter distal regions (Figure 3.2 C).

To validate these OCRs, we compared cell-type-specific peaks to DNase Hypersensitive sites (DHS) identified in bulk samples from ENCODE. OCRs from multiple cell types significantly overlapped with DHS from the left and right ventricles (40-60%; Figure 3.2D, top), however this overlap was mostly limited to cell-types with larger proportions. Compared to bulk data, snATAC-seq therefore added both cell-type-specificity and increased sensitivity for the detection of OCRs. We also observed strong overlap between cell-type-specific peaks and H3K27ac signal from LV and RV (60-80%), suggesting that these peaks are located within putative enhancer regions (Figure 3.2D, bottom).

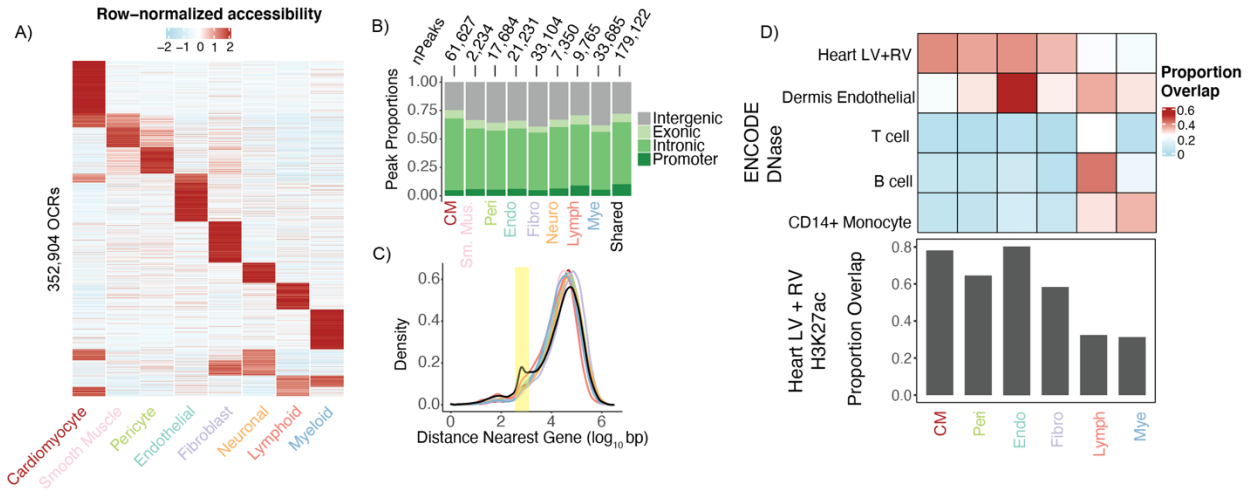

Figure 3.2: Discovery of open chromatin regions (OCRs) in the human heart. A) Row-normalized accessibility of PREs across all cell-types. B) Number of cell-type-specific and shared PREs and their genomic distribution. C) Density plot of the $\log_{10}$ distance to nearest gene for all cell-type-specific and shared PREs. The yellow strip highlights a region where the shared PREs deviates from the cell-type PREs. D) Proportion of cell-type specific PREs that overlap with DNase from cell-type in several tissue and primary cell-types (LV = left ventricle, RV = right ventricle). Below is the proportion of overlap with H3K27ac regions.

Cell-type-specific OCRs harbor binding sites for distinct sets of lineage-specific transcription factors (TFs) and thus can be used to identify sequence-specific determinants of TF activity. We identified 260 TF motifs that were enriched within cell-type-specific OCRs in at least one cell type at FDR < 1% (Figure 3.3 A). However, the degeneracy of TF motifs and the sharing of similar or identical binding preferences within TF families make it difficult to rely on motif enrichment alone to identify active TFs in a particular cell type. To mitigate this problem, we correlated motif accessibility with inferred gene expression using accessibility-derived gene scores across all cells for each of the TFs. This comparison yielded a set of 158 TFs with medium-to-high correlation ($|r| > 0.5$). This set included several known transcriptional regulators in CMs, including MEF2A (Figure 3.3 B, top), TBX5 (Figure 3.3B, bottom), and GATA4 as well as other cell types such as CEBPA in fibroblasts and SOX9 in endothelial cells. We thus generated a compendium of putative lineage-specific transcriptional regulators.

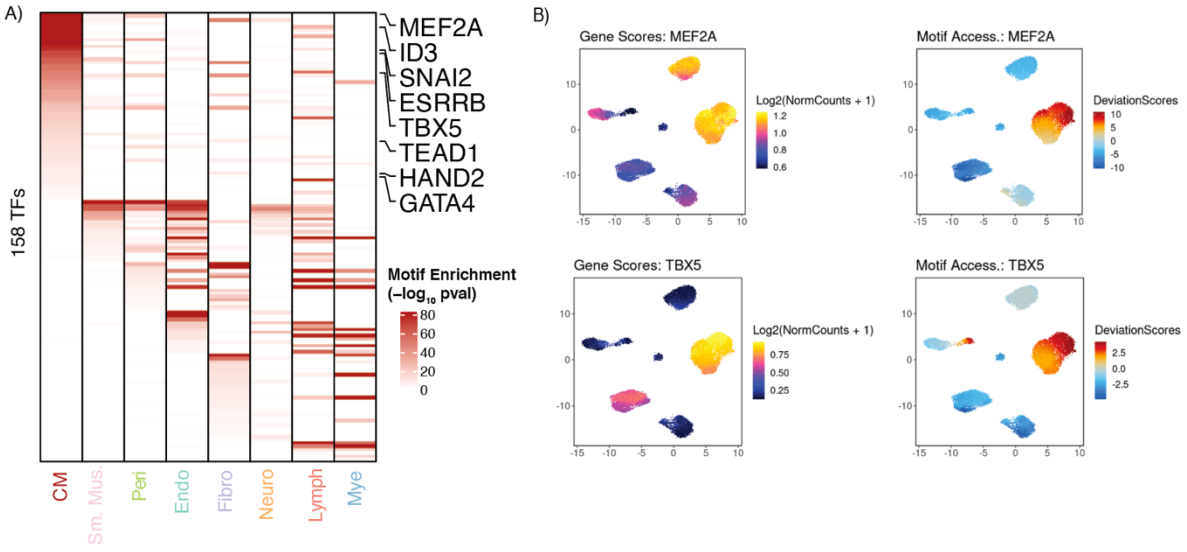

Figure 3.3: Transcription factor (TF) motif enrichment in OCRs of each cell-type. A) Enrichment heatmap of 260 TFs with at 1% FDR in at least one cell-type and an absolute gene score correlation of at least 0.5. B) Motif accessibility and gene score overlaid onto UMAP for two key TFs: MEF2A and TBX5.

Single-nucleus ATAC-seq also enables us to find putative interactions between distal enhancers and gene promoters through co-accessibility. Having these links will allows us later to link GWAS SNPs to putative target genes because most strong GWAS hits are in distal enhancers far away from gene bodies. To this end, we first calculated co-accessibility between all distal enhancer and gene promoter OCRs using CM cells only. Existing promoter-capture HiC (PC-HiC) data in iPSC-derived CMs[64] was used to benchmark our co-accessibility calculations. We find modest overlap between our co-accessible links and the links from PC-HiC data (Figure 3.4A). Having established the parameters for co-accessibility calculation, we estimated co-accessibility using meta-cells derived from all cells. Using these co-accessible links, we found on average 14 OCRs linked to any gene's promoter (Figure 3.4B). To further validate these co-accessible links, we subset the distal enhancers according to their cell-type specificity, and GSEA on the genes linked for each cell-type (Figure 3.4C). In doing this, we recover molecular function categories that are

characteristic of each cell-type, such as cardiac myofibril assembly for CMs and macrophage activation for myeloid cells. This compendium of putative enhancer-promoter interactions will be utilized, along with the PC-HiC data, in interpreting the genetics of AF.

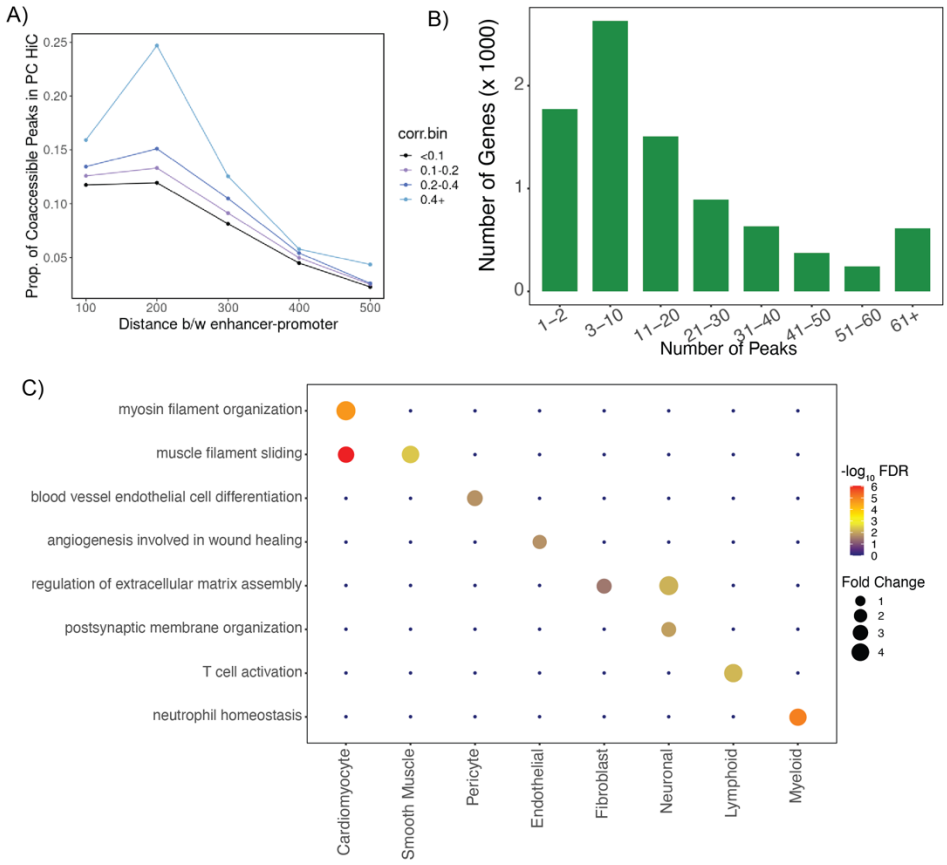

Figure 3.4: Linkage of distal enhancers to putative target genes. A) Proportion of co-accessible links in promoter-capture HiC (PC-HiC) data, partitioned by distance and correlation strength. B) Distribution of number of distal OCRs linked to gene promoters. C) Gene-set enrichment analysis of genes linked to cell-type specific OCRs via co-accessibility.

3.3.2 Fine-mapping atrial fibrillation GWAS with functional priors from snATAC-seq

A major challenge for the translation of GWAS findings into a fuller understanding of disease mechanisms is the identification of non-coding causal variants and the cell-types in which they are active. We assessed the enrichment of GWAS summary statistics for cardiac (AF, PR interval, heart rate), cardiovascular (CAD, blood pressure) and non-cardiovascular (BMI, height,

schizophrenia) traits using TORUS (Figure 3.5 A). We found strong CM-specific enrichment for AF and PR interval (>10-fold). We also found strong enrichment of endothelial/fibroblast/pericyte OCRs for coronary artery disease and blood pressure. Finally, we find no significant enrichment for non-cardiovascular traits, with the exception of height.

Given the strength of enrichment and cell-type specificity of CM OCRs in AF, we sought to perform fine-mapping on AF in order to infer causal variants and genes. Recently, a large scale GWAS on AF in a European population was performed where 111 signals/loci were observed at genome-wide significance, corresponding to 122 approximately independent LD blocks. As outlined in Section 3.2.3, we performed functionally-informed fine-mapping of all 122 blocks utilizing the union set of OCRs. We split the union set into CM-specific OCRs, CM-shared OCRs, and OCRs not in CMs. We also used coding and conserved annotations from UCSC. In this way, the entire union set is utilized, while SNPs in CM-specific/shared OCRs are prioritized. Under this framework, we find 54 SNPs with PIP > 50% (Table 3.2), compared with 39 SNPs in the case of no prior used (Figure 3.5 B). Visualizing the chromatin accessibility of the top 54 SNPs, we find a majority (31/54) are in CM-specific or shared OCRs as expected (Figure 3.5 C). We also find that 31/54 SNPs are in heart active enhancers as marked by H3K27ac. We also find 33/54 SNPs to be in enhancers that are co-accessible or have a PC-HiC link with a protein coding gene. Finally, we find 12/54 SNPs that have a strong TF motif disruption score for the cardiac 158 TFs previously defined (Figure 3.5 C). We also performed colocalization analysis with heart left ventricle eQTL data and found 5/54 SNPs to have eQTL effects in the heart. Altogether, these results support potential regulatory functions of a large fraction of fine-mapped variants, and implicate TF binding disruption as one main mechanism of functional effects of these variants.

Locus	SNP	Chr	b37 bp	REF	ALT	PIP	Gene Linked	Link Method	Gene PIP	Chromatin status	Distance to Gene
7	rs880315	1	10796866	C	T	0.754	CASZ1	Co-access.	0.34	CM Specific ATAC	59841
25	rs2885697	1	41544279	G	T	0.978	SLFNL1	Distance	0.403	Unannotated	55370
31	rs11590635	1	49309764	A	G	0.974	BEND5	Distance	0.949	Unannotated	67123
32	rs72692218	1	51436946	G	A	0.537	TTC39A	Co-access.	0.937	CM Shared ATAC	373842
72	rs4073778	1	116297758	C	A	0.7	CASQ2	Co-access.	0.214	CM Specific ATAC	13644
77	rs4999127	1	154714006	G	A	0.98	KCNN3	PC-HiC	0.437	Other	128750
78	rs11264280	1	154862952	T	C	0.999	KCNN3	PC-HiC	0.437	Unannotated	20196
85	rs72700118	1	170194823	A	C	0.579	PRRX1	Distance	0.602	Other	437046
86	rs577676	1	170587340	T	C	0.731	PRRX1	Distance	0.602	Other	44529
104	rs3737883	1	203034906	A	G	0.747	CHI3L1	PC-HiC	0.252	CM Specific ATAC	120971
187	rs72926475	2	86594487	A	G	0.809	IMMT	PC-HiC	0.341	CM Specific ATAC	171594
219	rs10496971	2	145769943	G	T	0.805	ZEB2	Co-access.	0.975	CM Shared ATAC	487796
238	rs35215597	2	175543782	G	A	0.84	WIPF1	Co-access.	0.364	CM Shared ATAC	3862
286	rs7650482	3	12841804	A	G	0.741	TMEM40	Co-access.	0.25	CM Specific ATAC	30848
304	rs116544863	3	38172474	T	C	0.837	SLC22A13	Co-access.	0.59	CM Specific ATAC	134829
305	rs6801957	3	38767315	T	C	0.829	SLC22A13	Co-access.	0.59	CM Specific ATAC	460012
397	rs60902112	3	194800853	T	C	0.556	TMEM44	Co-access.	0.937	CM Specific ATAC	446435
453	rs1458038	4	81164723	T	C	0.967	FGF5	Co-access.	0.981	CM Shared ATAC	23030
470	rs536594981	4	111106471	C	T	0.646	PITX2	Distance	0.944	Unannotated	456808
645	rs34969716	6	18210109	A	G	0.999	RNF144B	Co-access.	0.25	CM Specific ATAC	158670
660	rs3176326	6	36647289	A	G	1	RAB44	PC-HiC	0.125	CM Shared ATAC	35967
710	rs9481842	6	118974798	G	T	0.99	PLN	Distance	0.813	Unannotated	105337
728	rs117984853	6	149399100	T	G	1	GINM1	Co-access.	0.5	Non-CM ATAC	488330
791	rs74910854	7	74110705	G	A	0.809	GTF2I	Distance	0.366	Unannotated	38711
813	rs3807989	7	116186241	A	G	0.758	CAV1	Co-access.	0.324	CM Specific ATAC	21402
821	rs55985730	7	128417044	G	T	0.906	CALU	PC-HiC	0.467	Unannotated	37698
857	rs35620480	8	11499908	C	A	0.967	GATA4	Co-access.	0.249	CM Specific ATAC	34560
861	rs208757	8	17809791	G	A	0.909	PCM1	Distance	0.612	Unannotated	29442
862	rs7508	8	17913970	G	A	1	PCM1	Co-access.	0.612	CM Shared ATAC	133621
922	rs62521287	8	124552133	T	C	0.675	KLHL38	PC-HiC	0.19	CM Specific ATAC	113057
984	rs10821415	9	97713459	A	C	0.947	FBP2	Co-access.	0.485	CM Specific ATAC	357384
1009	rs2274115	9	139094773	A	G	0.693	LHX3	Distance	0.375	Unannotated	2182
1058	rs60632610	10	75415677	T	C	0.959	SYNPO2L	Exon	0.315	CM Specific ATAC	7884
1076	rs11598047	10	105342672	G	A	0.802	CALHM2	Distance	0.218	Unannotated	130012
1080	rs10749053	10	112576695	T	C	0.714	PDCD4	Distance	0.336	Unannotated	54870
1197	rs58747679	12	26348304	C	T	0.825	BHLHE41	Co-access.	0.356	CM Shared ATAC	70244
1201	rs2045172	12	32980161	T	G	0.841	PKP2	PC-HiC	0.537	CM Specific ATAC	69613
1222	rs71454237	12	70013415	A	G	0.877	LRRC10	Distance	0.29	Unannotated	8473
1264	rs9506925	13	23368943	T	C	0.76	FGF9	PC-HiC	0.937	Other	1123421
1331	rs11156751	14	32990437	C	T	0.713	AKAP6	Distance	0.755	Unannotated	191958
1332	rs73241997	14	35173775	T	C	0.771	CFL2	Distance	0.444	Unannotated	10254
1352	rs1152591	14	64680848	A	G	0.959	AKAP5	PC-HiC	0.148	CM Specific ATAC	251369
1357	rs3814866	14	73361021	C	A	0.587	DPF3	promoters	0.493	CM Shared ATAC	212
1404	rs117361082	15	57929969	G	A	0.547	GCOM1	Exon	0.333	CM Shared ATAC	45863
1414	rs7172038	15	73667255	G	T	0.959	NPTN	PC-HiC	0.388	CM Specific ATAC	259220
1416	rs12908004	15	80676925	G	A	0.998	ARNT2	Co-access.	0.363	Unannotated	19767
1431	rs140185678	16	2003016	A	G	1	HAGH	Co-access.	0.111	CM Specific ATAC	125821
1468	rs2106261	16	73051620	T	C	0.994	ZFHX3	Distance	0.7	Unannotated	41977
1490	rs78744936	17	7461343	A	G	0.525	YBX2	PC-HiC	0.073	CM Shared ATAC	263409
1506	rs1453559	17	38020419	T	C	0.519	IKZF3	Exon	0.327	CM Shared ATAC	22
1557	rs9953366	18	46474192	T	C	0.915	C18orf32	PC-HiC	0.278	CM Shared ATAC	539430
1670	rs2834618	21	36119111	G	T	1	CLIC6	Distance	0.483	Unannotated	77423
1682	rs464901	22	18597502	C	T	0.853	MICAL3	Co-access.	0.999	CM Specific ATAC	90177
1686	rs133902	22	26164079	T	C	0.609	MYO18B	Splice Junc.	0.985	Other	25968

Table 3.2: Prioritized AF SNPs with PIP > 50% and their RefSeq IDs, position, alleles, and the gene with highest gene PIP that is linked to each SNP.

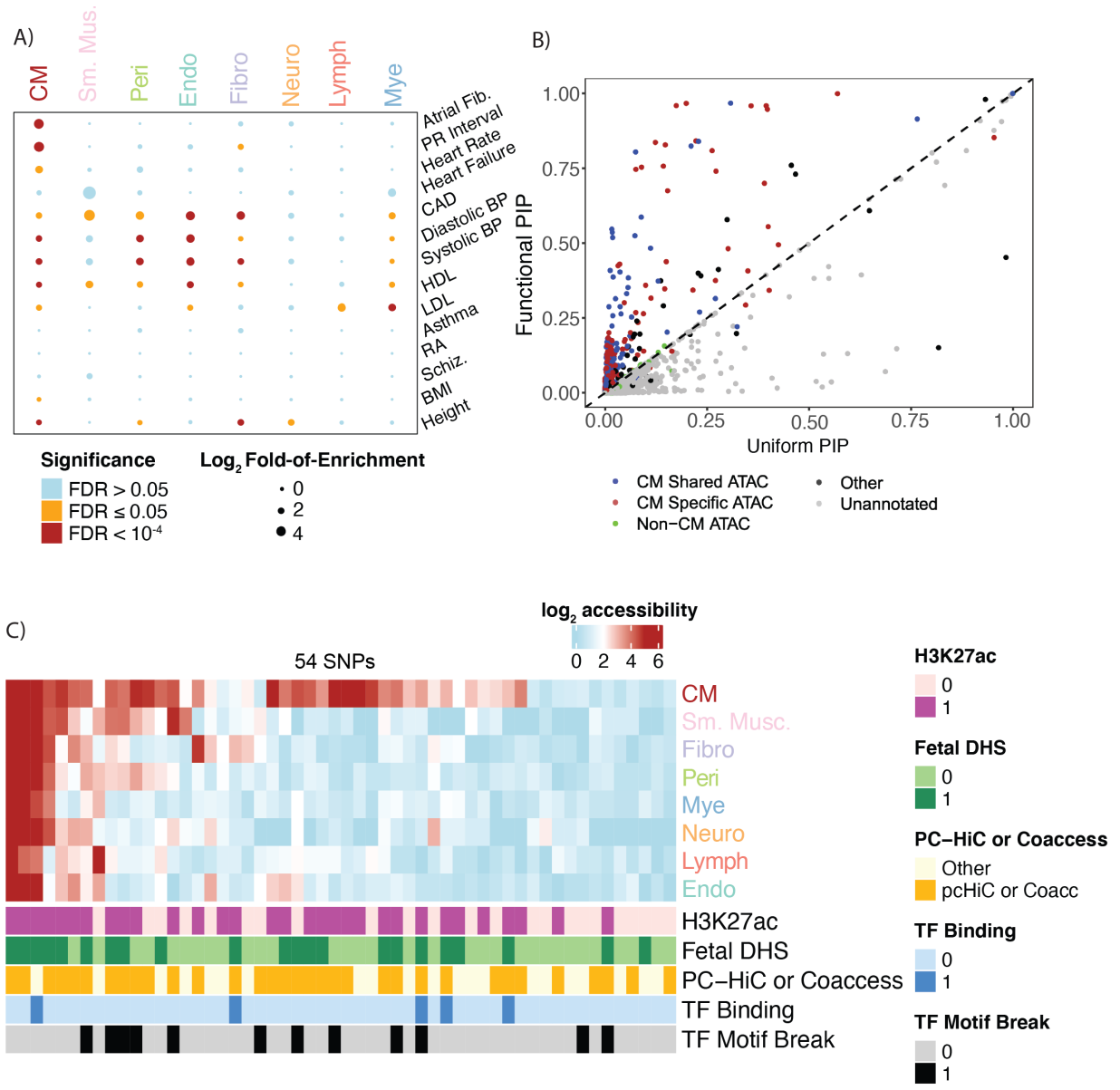

Figure 3.5: GWAS enrichment and fine-mapping summary of atrial fibrillation (AF). A) Fold of Enrichment and its significance for cell-type specific OCRs in each GWAS trait. B) Functional PIP vs uniform PIP from AF fine-mapping. C) Summary of AF SNPs with PIP > 50%. Top heatmap represents the $\log_2$-accessibility in a 500 bp window around each SNP in the specific OCRs of each cell-type. Bottom annotations represent binary annotations for each SNP: Heart left/right ventricle H3K27ac, fetal heart DHS, whether the SNP is linked via PC-HiC or co-accessibility (Coaccess), ChIP-seq regions for TBX5/GATA4/NKX2-5, and whether the SNP strongly disrupts cardiac TF motifs.

Most blocks were fine-mapped to at least two causal variants or more, with a mean number of 9 causal variants per locus (Figure 3.6 A). Therefore, the exact causal variant may still be ambiguous

and the molecular intermediates are not inferred from fine-mapping at the SNP level. To address this issue, we sought to generate a gene-level summary of the fine-mapping results, as outlined in section 3.2.4. We identified 24 high-confidence genes at gene PIP > 0.8, and 44 genes a relaxed PIP cutoff of 50% (Fig. B). At the locus level, the gene credible sets contain a single gene at 27 out of 122 blocks, and 2-5 genes at 69 blocks (Figure 3.6 B). Our candidates at PIP > 0.5, include many known AF risk genes, including TFs with known functions in cardiac development such as TBX5 and PITX2, ion channels such as HCN4, and other genes involved in muscle contractions such as TTN. Importantly, in 17/44 cases, the genes are not the closest genes of the top SNPs by PIPs, highlighting the importance of considering distal regulation. We compared our candidate genes at each locus with those from an earlier paper, which are based largely on distance, but also some functional information of genes. Overall, the two lists agree in a large fraction of blocks. In 46 blocks, our procedure found different top genes. Some of these novel candidates have plausible functions, such as ETV1, PRRX1 and FGF9. In some other cases where the earlier work picked a single candidate gene, our procedure has credible sets with multiple genes, reflecting the uncertainty of our knowledge of causal genes.

We systematically assessed the plausible role of the candidate genes. Using a set of control genes around the top PIP SNPs, we find that these prioritized genes are generally higher in expression in CMs (Figure 3.6 C). Comparing against the same control genes, we find that the prioritized genes are enriched among the differentially expressed genes (Figure 3.6 D). We tested enrichment of Gene Ontology (GO) terms in this gene set. This analysis reveals strong enrichment in Biological Processes including cardiac ventricle formation, peripheral nervous system development, and regulation of heart rate by cardiac conduction. (Figure 3.6 E)

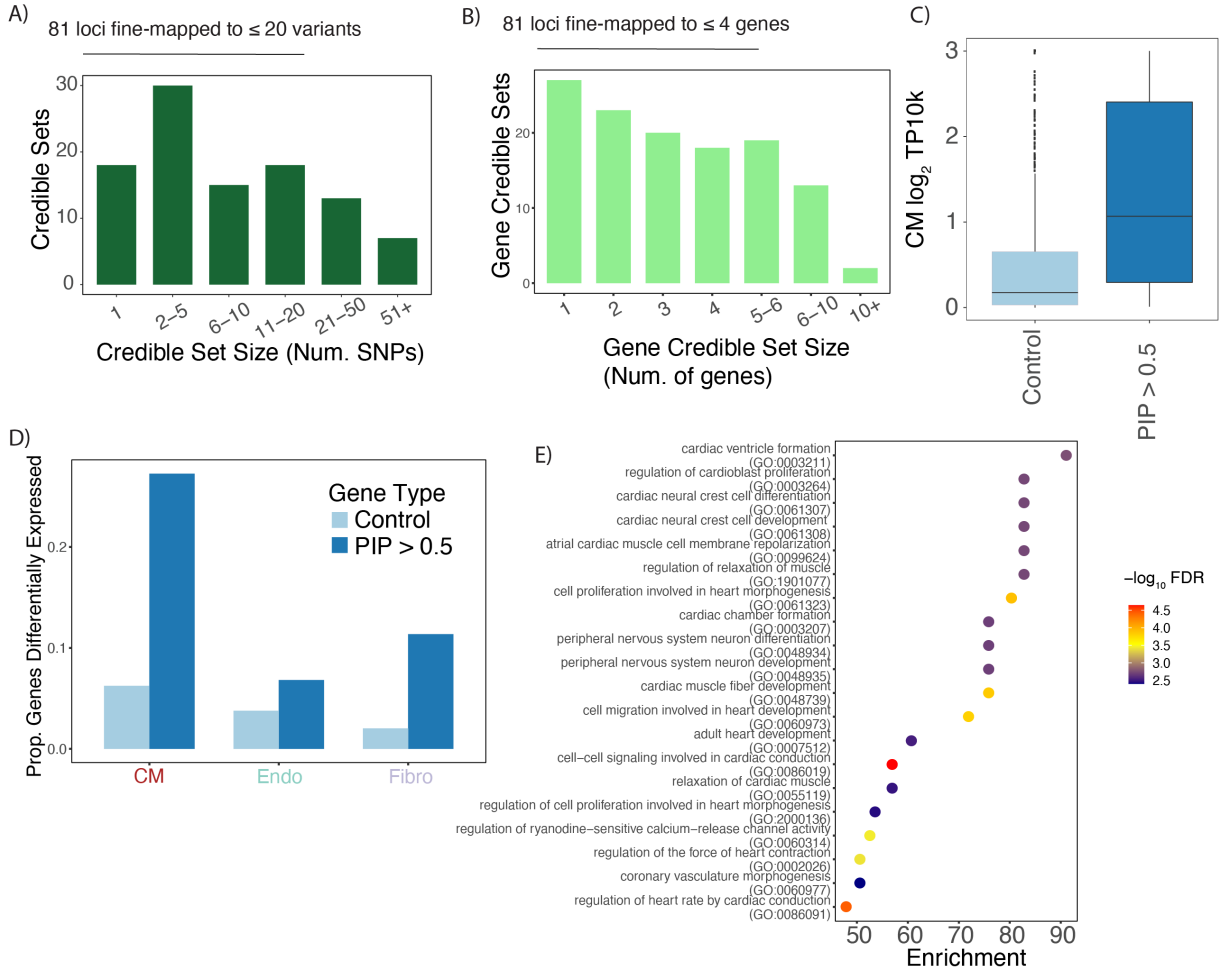

Figure 3.6: Gene fine-mapping summary. A) Number of SNPs in the credible set of each locus. Number of genes in the credible set of each locus C) Log_2 transcripts per 10k in cardiomyocyte (CM) pseudo-bulk for control genes and prioritized genes with PIP > 50%. D) Overlap of prioritized genes with differentially expressed genes, compared with control genes. E) Top 20 GO terms at FDR < 5% ordered by enrichment.

We now turn to two notable examples where our approach has yielded novel genes that highlight the strengths of our method. The first locus we zoom in on is T-box transcription factor 5 (TBX5), which is a key regulator of heart development and known to play a role in AF[65] (Figure 3.7, gene PIP = 0.95). An enhancer about 40kb upstream of the TBX5 promoter has >10 SNPs with strong LD and have above genome-wide significant association with AF. Using our functionally informed prior, we fine-mapped the region to a credible set with only 3 SNPs, with one SNP

(rs7312625) having a most of the signal (PIP = 0.4.) This SNP is in an enhancer that is active as marked by H3K27ac and is co-accessible with the promoter of TBX5.

A second notable example is the fibroblast growth factor 9 (FGF9), which has been shown to play a role in cardiomyocyte proliferation and differentiation[66] (Figure 3.8, gene PIP = 0.93). It is supported by two SNPs (PIP 0.6 and 0.2, respectively) that are linked, via PC-HiC, to the promoter of FGF9 nearly 1Mb away. Interestingly, the supporting SNPs of FGF9 lack regulatory marks in the adult heart, but are inside/close to fetal DHS, which interacts with the FGF9 promoter. This result thus suggests the possibility that some AF variants may act on the early developmental stage.

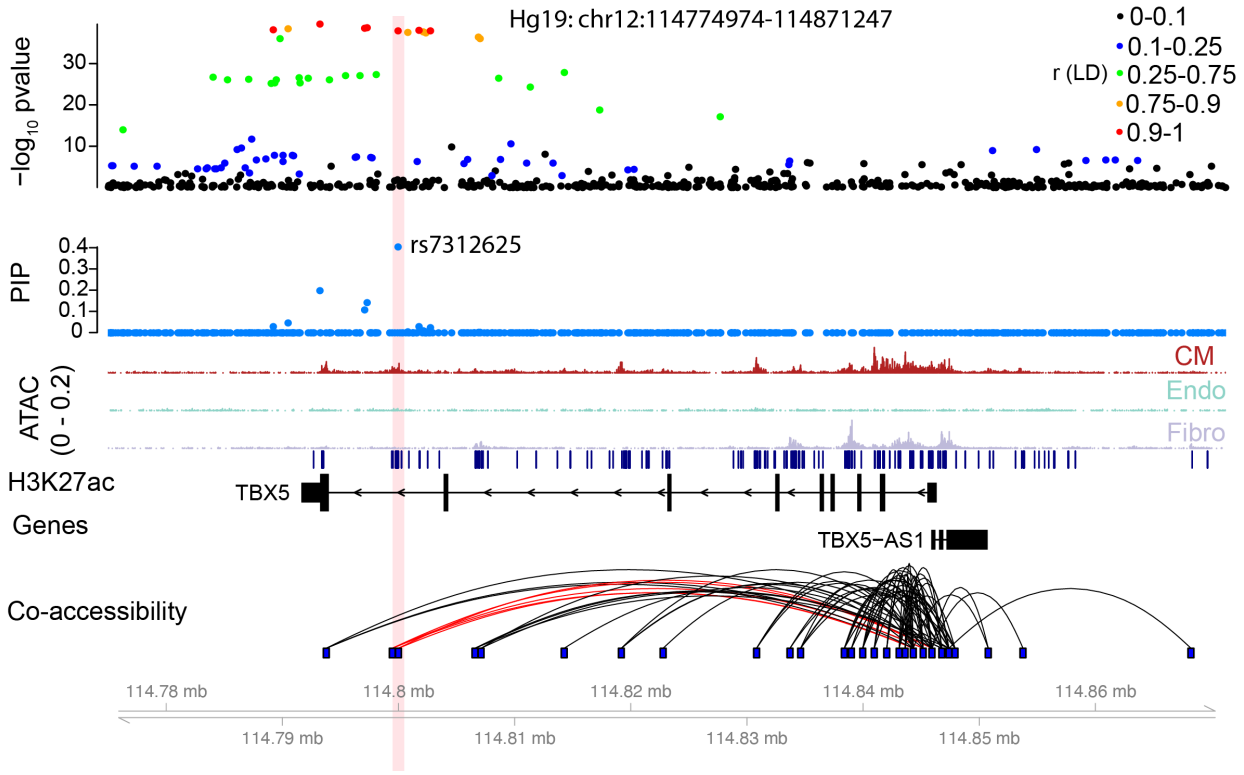

Figure 3.7: Locus track plot for T-box transcription factor 5 (TBX5).

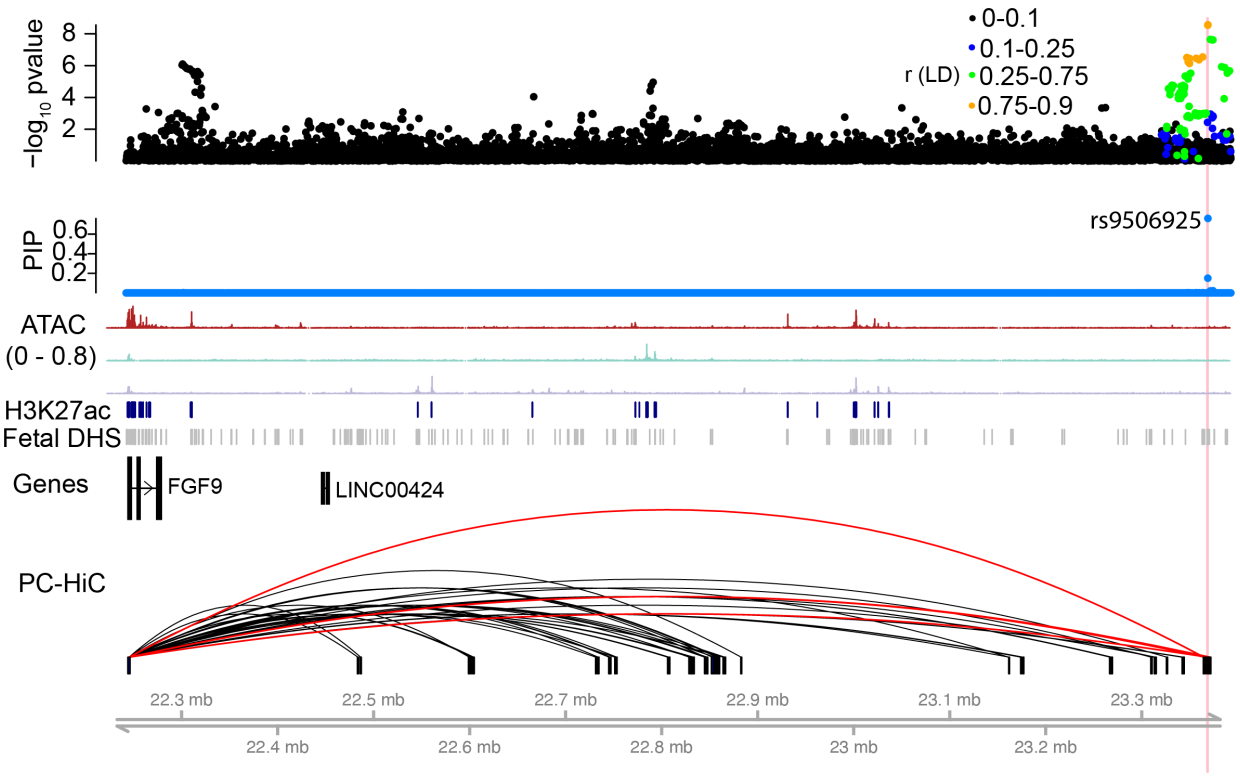

Figure 3.8: Locus track for fibroblast growth factor 9 (FGF9).

3.4 Discussion

We established a cell-type-resolved atlas of chromatin accessibility of the human heart and used this to identify regulatory factors and to shed light on the genetics complex traits. We used this to identify cell-type-specific sequence determinants associated with chromatin accessibility and inferred likely lineage-specific TFs.

We and others have previously incorporated functional annotations into fine-mapping procedures. However, these annotations were mostly derived from bulk samples or, in the case of immune cells, sorted cell-populations and thus did not represent cell-type-resolved information of all cell types present in the relevant organ. snATAC-seq provides a straightforward approach to obtain high-quality functional annotations from a single assay. In principle, the sensitive and accurate

identification of regulatory features should help to identify candidate regulatory regions. The strong enrichment of CM-specific regulatory features in AF SNPs suggest that a significant proportion of the AF risk localizes to regulatory features identified by snATAC-seq in ventricles.

A more general caveat to our analysis is that we attempted to apportion/explain disease association of variants in non-coding distal regions using chromatin accessibility of adult human hearts. It is possible that a subset of disease variants exerts its effects during development. Finally, we used a single assay, snATAC-seq, to define functional elements but it is likely that additional features of chromatin organization would refine annotations and improve fine-mapping. Future work might address these limitations by including samples from atria and by using additional single-cell genomics approaches, especially multimodal measurements are appealing because they represent a single assay solution.

We developed a novel computational strategy to identify causal genes. By considering each variant independently and linking it to genes based on genomic annotations, co-accessibility, PC-HiC, or distance, this 'gene fine-mapping' procedure accounts for uncertainty of causal variants and considers multiple ways a SNP can affect a gene.

A large fraction of the genes identified in this way are TFs, ion channels, and cardiac signaling proteins with known or plausible roles in heart development and/or AF. Furthermore, the 44 candidate genes show highly CM-specific expression pattern. We note that incorporation of the PC-HiC data in CM-iPSCs, which was publicly available, clearly helped this association. However, such auxiliary datasets are not available in most cases. Incorporation of additional measures (e.g. activity-by-contact scores[67]) or more robust establishment of co-accessibility measures are therefore desirable.

Our strategy presents a principled approach to identify and prioritize candidate causal genes including those that are distal to a SNP or separated by non-target genes. We provide a comprehensive resource of putative causal genes in AF. By narrowing the number of candidate genes per locus and providing gene PIPs, we support a rational procedure to select target genes for functional validation.

CHAPTER 4: CONCLUSION

Single-cell sequencing was the Method of the Year in 2013, and its application has only grown since due to commercialization of off-the-shelf equipment and analysis tools. In light of this, we have shifted our perspective on tissues and organs from homogenous blobs of cells to heterogenous cell mixtures with differing transcriptional and epigenetic programs. This has greatly aided our understanding of diseases, such as when scRNA-seq on human lung revealed an ionocyte cell type that differentially expresses CFTCR and may be causal for cystic fibrosis[68]. In other developmental systems such as the zebrafish embryo, a new cell state was inferred that has the hallmarks of apoptosis and cellular stress[68]. Progress from single-cell sequencing has been particularly remarkable in the nervous and immune systems, where dozens or even hundreds of transcriptionally distinct cell types have been mapped. Thanks to these efforts, we know that many diseases are driven by a select few cell-types with specialized transcriptional programs in an organ.

On the other hand, variation in DNA has been extensively studied and linked to disease phenotype. For example, GWAS has identified hundreds of loci for individual traits that are strongly associated with disease phenotypes. Given that every cell in an organism contains the same DNA, it becomes quite the challenge to pin down why a certain DNA variant is associated with the phenotype. In certain diseases, such as Type-2 diabetes (T2D), it is clear that the insulin-secreting beta cell of pancreatic islets is one of the causal cell types where T2D SNPs exert their effects. In other diseases, such as coronary artery disease, the germline SNPs associated likely exert their effects in multiple cell types, such as endothelial cells, fibroblasts, and pericytes/smooth muscle cells. This intuition of where a disease takes place in the human body has driven researchers to use bulk tissue assays on the relevant organ to further interpret germline SNPs. However, bulk assays

limit our understanding of which cell-type(s) or cell-state(s) are driving the disease phenotype. In this dissertation, I have developed a framework to utilize single-cell 'omics data to interpret DNA variants associated with disease states. This interpretation involves learning the relevant cell-type(s) for the disease, and the cell-type specific genes, e.g. the gene product, that mediates the effect of the SNP onto the phenotype.

In Chapter 1, I estimated genotypes of somatic point mutations in individual breast cancer cells, and performed evolution-informed clustering of cells. In doing this, I found two groups (early and late) of cancer cells with differing somatic mutational profiles. Performing differential gene expression between these two groups, the late cells showed a marked expression of breast cancer oncogenes and down-regulation of tumor suppressor genes. The set of somatic mutations that are unique to the late cells are putatively associated with the set of genes that are differentially expressed. Given that the method requires relatively high sequencing depth per cell, we were only able to perform this genetic phylogeny reconstruction on 18 cells. We likely underestimate the genetic heterogeneity given the small number of cells, and future efforts should focus on multi-regional sampling from tissues to better sample the genetic heterogeneity in tumors. High through-put scRNA-seq can sample an order of magnitude more cells at the cost of sequencing depth. While our method would likely struggle on such data, an alternative approach utilizing somatic copy number variants requires less sequencing depth and may provide similar single-cell phylogenies.

We have observed in Chapter 1 that scRNA-seq, paired with WES allows us to obtain higher resolution phylogenies with additional information such as transcriptome readout. However, obtaining single cells from tissues that are hard to dissociate, composed of fragile cells, and/or frozen specimens is not always possible. As an alternative, DroNc-seq, a high-throughput single-nucleus RNA sequencing protocol, has the potential to reveal tissue heterogeneity, at scale, based

on nuclear RNA, and is being increasingly used to profile primary tissue at high throughput. In Chapter 2, I investigated the performance of single-cell and single-nucleus RNA-sequencing in determining cell-types in a dynamic biological system. Single-cell generally detected more UMI and genes per cell, as expected given the higher amount of RNA material from whole cell. One major finding is that nearly half of reads from DroNc-seq were intronic reads, possibly stemming from internal priming on the introns of nascent RNAs. We incorporated these intronic reads for quantifying gene expression and we observe a 50% increase in gene detection rate. We found that for cell-type quantification at a fairly course-grain level, both methods are able to detect the same cellular populations. Increasing the cell-type resolution to define finer-grain communities will likely result in differences between Drop-seq and DroNc-seq. In conclusion, we find DroNc-seq as a viable alternative for measuring expression profiles from complex and heterogenous banked tissues and organs.

In Chapter 3, I analyzed the cellular heterogeneity found in adult human heart tissue using single-nucleus RNA and ATAC sequencing data. These methods defined the transcriptional and open chromatin landscape for eight major cells, such as cardiomyocytes, cardiac fibroblasts, and an immune compartment. We used these data to interpret the genetic architecture of cardiac diseases. The strong enrichment of CM-specific regulatory features in AF SNPs suggest that a significant proportion of the AF risk localizes to regulatory features identified by snATAC-seq in heart ventricles. Indeed, CMs are the key cell-type for which AF SNPs exert their effects through. To ascertain the causal variants, we performed statistical fine-mapping with CM open chromatin regions as a functional prior. We found 54 high-confidence SNPs across 122 blocks, with an average credible-set size of 9 SNPs per locus. More than half of these SNPs were in open chromatin regions in CMs while also overlapping with heart H3K27ac regions, therefore we likely prioritized

functional variants that have enhancer activity. Furthermore, we find >7 fold enrichment of these SNPs in the binding sites of TBX5, NKX2-5, and GATA4, suggesting the importance of these TFs. Despite fine-mapping, some blocks still contained multiple potentially causal variants, making interpretation difficult. Therefore, we aggregated the signal from fine-mapped SNPs onto genes using a novel computational approach. This approach may be better than assigning top SNPs to the nearest genes for two reasons. One, long-range transcriptional regulation by distal enhancers allows for a regulatory variant to modulate the gene activity from up to 1Mb away. These long-range interactions can be measured by HiC and other derivatives and we utilize such data to link SNPs to their potential target genes. Second, the signal from multiple SNPs with weak effects can be aggregated, leading to a strong signal at the gene level. For example, the gene CAMK2D has a gene PIP of >95% due to the cumulative signal of 8 SNPs with diffused PIPs. Such an example reflects that uncertain fine-mapping at the SNP-level can sometimes lead to high certainty at the gene-level.

We found 44 high confidence genes, some of which would not be implicated by GWAS alone. These genes allowed us to further interpret the genetics behind AF. In addition to calcium and potassium ion channels that are involved in cardiac conduction, we found multiple genes that are TFs, many of which have been implicated in heart development such as TBX5[65], PITX2[69], and HAND2[70]. Indeed, the most significant GO terms for the genes we nominate are involved in heart cell differentiation and proliferation. The over-representation of TFs is partly due to our utilization of long-range loops from PC-HiC for linking SNPs to genes. The presence of well-known cardiac TFs among the AF candidate genes and enrichment of their binding sites at putative causal SNPs support the importance of a TF-centered gene regulatory network mediating the AF

risk. I look forward to further experimental validation of the nominated AF enhancers and genes through the use of realistic cell lines and/or mouse models.

While our gene-mapping approach seems to work well for AF, there are potential issues and limitations in its application to other traits. The procedure relies on high quality SNP level fine-mapping results, which may not be the case for other low-powered traits. AF genetics are mostly enriched in cardiomyocytes only, which partially overcomes ambiguity from LD leading to many high PIP SNPs. This may not be the case for other traits that are more complex in terms of cell-type, such coronary artery disease as we saw in our enrichment analysis.

There are three major improvements that can be made to the general framework of linking germline SNPs to their putative target genes. A recent model called activity-by-contact (ABC)[67] has been proposed for linking SNPs to genes. The ABC score is obtained by taking the product of the regulatory activity of the sequence in which the variant resides in (using chromatin accessibility or similar) and the strength of the physical contact with a gene promoter from HiC data. Currently, a variant is assigned to a gene via PC-HiC in a binary fashion, therefore we may miss many possible SNP-gene interactions. A second possible improvement is the use of other technologies that better capture long-range chromatin interactions. For instance, chromatin interaction analysis with paired-end tag (ChIA-PET)[71] has been shown to have higher resolution than HiC. Furthermore, ChIA-PET can be used to construct chromatin interaction networks bound by proteins such as histones with H3K27ac, therefore yielding interactions that only involve active enhancers, leading to improved gene fine-mapping. Finally, a GWAS locus may have multiple causal signals which we do not consider in our current gene-mapping approach. GWAS loci often harbor multiple, independent causal variants, potentially all targeting the same genes (known as allelic heterogeneity, or AH). For example, the TCF7L2 locus of T2D contains at least 8

independent signals[72]. Indeed, aggregating multiple causal signals that target the same gene can potentially improve gene-mapping.

REFERENCES

[1] B. Alberts, “Molecular Biology of the Cell - NCBI Bookshelf.” https://www.ncbi.nlm.nih.gov/books/NBK21054/ (accessed Jun. 20, 2021).

[2] C. Greenman *et al.*, “Patterns of somatic mutation in human cancer genomes,” *Nature*, vol. 446, no. 7132, pp. 153–158, Mar. 2007, doi: 10.1038/nature05610.

[3] A. Buniello *et al.*, “The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019,” *Nucleic Acids Res.*, vol. 47, no. D1, pp. D1005–D1012, Jan. 2019, doi: 10.1093/nar/gky1120.

[4] R. E. Thurman *et al.*, “The accessible chromatin landscape of the human genome,” *Nature*, vol. 489, no. 7414, pp. 75–82, Sep. 2012, doi: 10.1038/nature11232.

[5] E. Z. Macosko *et al.*, “Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets,” *Cell*, vol. 161, no. 5, pp. 1202–1214, 2015, doi: 10.1016/j.cell.2015.05.002.

[6] A. Poran *et al.*, “Single-cell RNA sequencing reveals a signature of sexual commitment in malaria parasites,” *Nature*, 2017, doi: 10.1038/nature24280.

[7] N. Karaiskos *et al.*, “The Drosophila embryo at single-cell transcriptome resolution,” *Science (80-.).*, 2017, doi: 10.1126/science.aan3235.

[8] J. Qian *et al.*, “A pan-cancer blueprint of the heterogeneous tumor microenvironment revealed by single-cell profiling,” *Cell Res.*, vol. 30, no. 9, pp. 745–762, Sep. 2020, doi: 10.1038/s41422-020-0355-0.

[9] J. Lee, D. Y. Hyeon, and D. Hwang, “Single-cell multiomics: technologies and data analysis methods,” *Experimental and Molecular Medicine*, vol. 52, no. 9. Springer Nature, pp. 1428–1442, Sep. 01, 2020, doi: 10.1038/s12276-020-0420-2.

[10] S. Bian *et al.*, “Single-cell multiomics sequencing and analyses of human colorectal cancer,” *Science (80-.).*, vol. 362, no. 6418, pp. 1060–1063, Nov. 2018, doi: 10.1126/science.aao3791.

[11] K. Jahn, J. Kuipers, and N. Beerenwinkel, “Tree inference for single-cell data,” *Genome Biol.*, vol. 17, no. 1, p. 86, May 2016, doi: 10.1186/s13059-016-0936-x.

[12] N. Habib *et al.*, “Massively parallel single-nucleus RNA-seq with DroNc-seq,” *Nat. Methods*, vol. 14, no. 10, pp. 955–958, 2017, doi: 10.1038/nmeth.4407.

[13] J. B. Nielsen *et al.*, “Biobank-driven genomic discovery yields new insight into atrial fibrillation biology,” *Nature Genetics*, vol. 50, no. 9. Nature Publishing Group, pp. 1234–

1239, Sep. 01, 2018, doi: 10.1038/s41588-018-0171-3.

[14] G. Wang, A. Sarkar, P. Carbonetto, and M. Stephens, "A simple new approach to variable selection in regression, with application to genetic fine mapping," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 82, no. 5, pp. 1273–1300, Dec. 2020, doi: 10.1111/rssb.12388.

[15] C. I. Wu, H. Y. Wang, S. Ling, and X. Lu, "The Ecology and Evolution of Cancer: The Ultra-Microevolutionary Process," *Annual Review of Genetics*, vol. 50. Annual Reviews Inc., pp. 347–369, Nov. 23, 2016, doi: 10.1146/annurev-genet-112414-054842.

[16] S. Nik-Zainal *et al.*, "The life history of 21 breast cancers," *Cell*, vol. 149, no. 5, pp. 994–1007, May 2012, doi: 10.1016/j.cell.2012.04.023.

[17] L. R. Yates and P. J. Campbell, "Evolution of the cancer genome," *Nature Reviews Genetics*, vol. 13, no. 11. Nature Publishing Group, pp. 795–806, Nov. 09, 2012, doi: 10.1038/nrg3317.

[18] M. Greaves and C. C. Maley, "Clonal evolution in cancer," *Nature*, vol. 481, no. 7381. Nature Publishing Group, pp. 306–313, Jan. 19, 2012, doi: 10.1038/nature10762.

[19] S. Vosberg and P. A. Greif, "Clonal evolution of acute myeloid leukemia from diagnosis to relapse," *Genes Chromosom. Cancer*, vol. 58, no. 12, pp. 839–849, Dec. 2019, doi: 10.1002/gcc.22806.

[20] R. J. Gillies, D. Verduzco, and R. A. Gatenby, "Evolutionary dynamics of carcinogenesis and why targeted therapy does not work," *Nature Reviews Cancer*, vol. 12, no. 7. Nat Rev Cancer, pp. 487–493, Jul. 2012, doi: 10.1038/nrc3298.

[21] N. E. Navin, "Cancer genomics: one cell at a time," *Genome biology*, vol. 15, no. 8. BioMed Central, p. 452, Aug. 30, 2014, doi: 10.1186/s13059-014-0452-9.

[22] Y. Qiao, A. R. Quinlan, A. A. Jazaeri, R. G. w. Verhaak, D. A. Wheeler, and G. T. Marth, "SubcloneSeeker: a computational framework for reconstructing tumor clone structure for cancer variant interpretation and prioritization," *Genome Biol.*, vol. 15, no. 8, p. 443, Aug. 2014, doi: 10.1186/s13059-014-0443-x.

[23] A. G. Deshwar, S. Vembu, C. K. Yung, G. H. Jang, L. Stein, and Q. Morris, "PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors," *Genome Biol.*, vol. 16, no. 1, p. 35, Feb. 2015, doi: 10.1186/s13059-015-0602-8.

[24] F. Strino, F. Parisi, M. Micsinai, and Y. Kluger, "TrAp: A tree approach for fingerprinting subclonal tumor composition," *Nucleic Acids Res.*, vol. 41, no. 17, Sep. 2013, doi: 10.1093/nar/gkt641.

[25] W. Chung *et al.*, "Single-cell RNA-seq enables comprehensive tumour and immune cell

profiling in primary breast cancer," *Nat. Commun.*, vol. 8, no. 1, pp. 1–12, May 2017, doi: 10.1038/ncomms15081.

[26] Q. Sheng, S. Zhao, C. I. Li, Y. Shyr, and Y. Guo, "Practicability of detecting somatic point mutation from RNA high throughput sequencing data," *Genomics*, vol. 107, no. 5, pp. 163–169, May 2016, doi: 10.1016/j.ygeno.2016.03.006.

[27] G. A. Van der Auwera *et al.*, "From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline," *Curr. Protoc. Bioinforma.*, vol. 43, no. SUPL.43, pp. 11.10.1-11.10.33, Oct. 2013, doi: 10.1002/0471250953.bi1110s43.

[28] A. Dobin *et al.*, "STAR: Ultrafast universal RNA-seq aligner," *Bioinformatics*, 2013, doi: 10.1093/bioinformatics/bts635.

[29] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, Jul. 2009, doi: 10.1093/bioinformatics/btp324.

[30] J. Yang, W. Liu, X. Lu, Y. Fu, L. Li, and Y. Luo, "High expression of small GTPase Rab3D promotes cancer progression and metastasis," *Oncotarget*, vol. 6, no. 13, pp. 11125–11138, 2015, doi: 10.18632/oncotarget.3575.

[31] O. Rozenblatt-Rosen, M. J. T. Stubbington, A. Regev, and S. A. Teichmann, "The Human Cell Atlas: From vision to reality," *Nature*. 2017, doi: 10.1038/550451a.

[32] A. M. Klein *et al.*, "Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells Accession Numbers GSE65525 Klein et al Resource Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells," *Cell*, 2015, doi: 10.1016/j.cell.2015.04.044.

[33] B. B. Lake *et al.*, "A comparative strategy for single-nucleus and single-cell transcriptomes confirms accuracy in predicted cell-type expression from nuclear RNA," *Sci. Rep.*, 2017, doi: 10.1038/s41598-017-04426-w.

[34] T. E. Bakken *et al.*, "Single-nucleus and single-cell transcriptomes compared in matched cortical cell types," *PLoS One*, 2018, doi: 10.1371/journal.pone.0209648.

[35] H. Wu, Y. Kirita, E. L. Donnelly, and B. D. Humphreys, "Advantages of Single-Nucleus over Single-Cell RNA Sequencing of Adult Kidney: Rare Cell Types and Novel Cell States Revealed in Fibrosis," *J. Am. Soc. Nephrol.*, 2018, doi: 10.1681/asn.2018090912.

[36] N. E. Banovich *et al.*, "Impact of regulatory variation across human iPSCs and differentiated cells," *Genome Res.*, vol. 28, pp. 1243–1252, 2017.

[37] J. Köster and S. Rahmann, "Snakemake-a scalable bioinformatics workflow engine,"

Bioinformatics, 2012, doi: 10.1093/bioinformatics/bts480.

[38] S. Andrews and Babraham Bioinformatics, "FastQC: A quality control tool for high throughput sequence data," *Manual*. 2010, doi: citeulike-article-id:11583827.

[39] T. Smith, A. Heger, and I. Sudbery, "UMI-tools: Modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy," *Genome Res.*, 2017, doi: 10.1101/gr.209601.116.

[40] M. Martin, "Cutadapt removes adapter sequences from high-throughput sequencing reads," *EMBnet.journal*, 2011, doi: 10.14806/ej.17.1.200.

[41] Y. Liao, G. K. Smyth, and W. Shi, "FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features," *Bioinformatics*, 2014, doi: 10.1093/bioinformatics/btt656.

[42] A. Butler, P. Hoffman, P. Smibert, E. Papalexi, and R. Satija, "Integrating single-cell transcriptomic data across different conditions, technologies, and species," *Nat. Biotechnol.*, 2018, doi: 10.1038/nbt.4096.

[43] S. Freytag, L. Tian, I. Lönnstedt, M. Ng, and M. Bahlo, "Comparison of clustering tools in R for medium-sized 10x Genomics single-cell RNA-sequencing data," *F1000Research*, vol. 7, p. 1297, 2018, doi: 10.12688/f1000research.15809.2.

[44] V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech. Theory Exp.*, vol. 2008, no. 10, p. P10008, Oct. 2008, doi: 10.1088/1742-5468/2008/10/P10008.

[45] V. Y. Kiselev *et al.*, "SC3: Consensus clustering of single-cell RNA-seq data," *Nat. Methods*, 2017, doi: 10.1038/nmeth.4236.

[46] L. McInnes, J. Healy, N. Saul, and L. Großberger, "UMAP: Uniform Manifold Approximation and Projection," *J. Open Source Softw.*, 2018, doi: 10.21105/joss.00861.

[47] B. Etienne, D. Charles-Antoine, K. Immanuel W.H., N. Lai Guan, G. Florent, and N. Evan W., "Evaluation of UMAP as an alternative to t-SNE for single-cell data," *Development*, 2018, doi: 10.1101/298430.

[48] C. Soneson and M. D. Robinson, "Bias, robustness and scalability in single-cell differential expression analysis," *Nat. Methods*, 2018, doi: 10.1038/nmeth.4612.

[49] G. La Manno *et al.*, "RNA velocity of single cells," *Nature*. 2018, doi: 10.1038/s41586-018-0414-6.

[50] E. Mereu *et al.*, "Benchmarking Single-Cell RNA Sequencing Protocols for Cell Atlas

Projects," *bioRxiv*, 2019, doi: 10.1101/630087.

[51] C. E. Friedman *et al.*, "Single-Cell Transcriptomic Analysis of Cardiac Differentiation from Human PSCs Reveals HOPX-Dependent Cardiomyocyte Maturation," *Cell Stem Cell*, 2018, doi: 10.1016/j.stem.2018.09.009.

[52] L. W and V. RS, "Genetics of coronary artery disease," *Circulation*, vol. 128, no. 10, pp. 1131–1138, Sep. 2013, doi: 10.1161/CIRCULATIONAHA.113.005350.

[53] M. T. Maurano *et al.*, "Systematic Localization of Common Disease-Associated Variation in Regulatory DNA," *Science*, vol. 337, no. 6099, p. 1190, Sep. 2012, doi: 10.1126/SCIENCE.1222794.

[54] A. Mahajan *et al.*, "Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps," *Nat. Genet. 2018 5011*, vol. 50, no. 11, pp. 1505–1513, Oct. 2018, doi: 10.1038/s41588-018-0241-6.

[55] C. S. McGinnis, L. M. Murrow, and Z. J. Gartner, "DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors," *Cell Syst.*, vol. 8, no. 4, pp. 329-337.e4, Apr. 2019, doi: 10.1016/J.CELS.2019.03.003.

[56] J. M. Granja *et al.*, "ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis," *Nat. Genet. 2021 533*, vol. 53, no. 3, pp. 403–411, Feb. 2021, doi: 10.1038/s41588-021-00790-6.

[57] R. Fang *et al.*, "Comprehensive analysis of single cell ATAC-seq data with SnapATAC," *Nat. Commun. 2021 121*, vol. 12, no. 1, pp. 1–15, Feb. 2021, doi: 10.1038/s41467-021-21583-9.

[58] T. Stuart, A. Srivastava, C. Lareau, and R. Satija, "Multimodal single-cell chromatin analysis with Signac," *bioRxiv*, p. 2020.11.09.373613, Nov. 2020, doi: 10.1101/2020.11.09.373613.

[59] I. Korsunsky *et al.*, "Fast, sensitive and accurate integration of single-cell data with Harmony," *Nat. Methods 2019 1612*, vol. 16, no. 12, pp. 1289–1296, Nov. 2019, doi: 10.1038/s41592-019-0619-0.

[60] W. MT *et al.*, "Determination and inference of eukaryotic transcription factor sequence specificity," *Cell*, vol. 158, no. 6, pp. 1431–1443, 2014, doi: 10.1016/J.CELL.2014.08.009.

[61] B. T and P. JK, "Approximately independent linkage disequilibrium blocks in human populations," *Bioinformatics*, vol. 32, no. 2, pp. 283–285, Jan. 2016, doi: 10.1093/BIOINFORMATICS/BTV546.

[62] A. Auton *et al.*, “A global reference for human genetic variation,” *Nat. 2015 5267571*, vol. 526, no. 7571, pp. 68–74, Sep. 2015, doi: 10.1038/nature15393.

[63] X. Wen, “Molecular QTL discovery incorporating genomic annotations using Bayesian false discovery rate control,” *https://doi.org/10.1214/16-AOAS952*, vol. 10, no. 3, pp. 1619–1638, Sep. 2016, doi: 10.1214/16-AOAS952.

[64] L. E. Montefiori *et al.*, “A promoter interaction map for cardiovascular disease genetics,” *Elife*, vol. 7, Jul. 2018, doi: 10.7554/ELIFE.35788.

[65] G. DF *et al.*, “TBX5 loss-of-function mutation contributes to atrial fibrillation and atypical Holt-Oram syndrome,” *Mol. Med. Rep.*, vol. 13, no. 5, pp. 4349–4356, May 2016, doi: 10.3892/MMR.2016.5043.

[66] I. N, O. H, N. Y, and K. M, “Roles of FGF Signals in Heart Development, Health, and Disease,” *Front. cell Dev. Biol.*, vol. 4, no. OCT, Oct. 2016, doi: 10.3389/FCELL.2016.00110.

[67] C. P. Fulco *et al.*, “Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations,” *Nat. Genet. 2019 5112*, vol. 51, no. 12, pp. 1664–1669, Nov. 2019, doi: 10.1038/s41588-019-0538-0.

[68] A. F. Schier, “Single-cell biology: beyond the sum of its parts,” *Nat. Methods 2020 171*, vol. 17, no. 1, pp. 17–20, Jan. 2020, doi: 10.1038/s41592-019-0693-3.

[69] D. Franco, D. Sedmera, and E. Lozano-Velasco, “Multiple Roles of Pitx2 in Cardiac Development and Disease,” *J. Cardiovasc. Dev. Dis.*, vol. 4, no. 4, p. 16, Oct. 2017, doi: 10.3390/JCDD4040016.

[70] R. M. George and A. B. Firulli, “Hand Factors in Cardiac Development,” *Anat. Rec. (Hoboken).*, vol. 302, no. 1, p. 101, Jan. 2019, doi: 10.1002/AR.23910.

[71] G. Li *et al.*, “Chromatin Interaction Analysis with Paired-End Tag (ChIA-PET) sequencing technology and application,” *BMC Genomics 2014 1512*, vol. 15, no. 12, pp. 1–10, Dec. 2014, doi: 10.1186/1471-2164-15-S12-S11.

[72] S. F. A. Grant, “The TCF7L2 Locus: A Genetic Window Into the Pathogenesis of Type 1 and Type 2 Diabetes,” *Diabetes Care*, vol. 42, no. 9, p. 1624, Sep. 2019, doi: 10.2337/DCI19-0001.