UC Santa Cruz

UC Santa Cruz Electronic Theses and Dissertations

Title

Integrative Analysis of Single-Cell RNA and Single-Cell ATAC Profiles to Infer Gene Regulatory Networks

Permalink

https://escholarship.org/uc/item/1j9843tz

Author

Safavi, Rojin

Publication Date

2022

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA
SANTA CRUZ

INTEGRATIVE ANALYSIS OF SINGLE-CELL RNA AND SINGLE-CELL ATAC PROFILES TO INFER GENE REGULATORY NETWORKS

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

BIOMOLECULAR ENGINEERING AND BIOINFORMATICS

by

Rojin Safavi

September 2022

The Dissertation of Rojin Safavi
is approved:

Professor Josh Stuart, Chair

Professor Neil Chi

Professor Ali Shariati

Peter F. Biehl
Vice Provost and Dean of Graduate Studies

Copyright © by

Rojin Safavi

2022

Table of Contents

List of Figures **v**

List of Tables **vi**

Abstract **vii**

Dedication **x**

1 sci-GRN: Integrative Gene Regulatory Network Inference **1**
1.1 Abstract . . . 1
1.2 Introduction . . . 2
1.3 Background . . . 5
1.3.1 RNA sequencing reveals gene regulatory networks in development and disease . . . 5
1.3.2 Discovering transcription factor motifs reveals regulatory linkages 7
1.3.3 Single-cell sequencing . . . 9
1.3.4 Gene regulatory inference . . . 10
1.4 Methods . . . 15
1.4.1 Heart tissue samples . . . 15
1.4.2 scRNA-seq processing . . . 15
1.4.3 scATAC-seq processing . . . 16
1.4.4 scRNA and scATAC integration . . . 17
1.4.5 Gene regulatory network base . . . 19
1.4.6 Putative target gene ranking . . . 21
1.4.7 PBMC benchmarking . . . 25
1.5 Results . . . 37
1.5.1 Single-cell heart development dataset . . . 37
1.5.2 Single-cell RNA and single-cell ATAC integration . . . 40
1.5.3 Detecting enhancer dependent regulations . . . 45
1.5.4 TBX5 target genes . . . 47
1.5.5 Regulon activity inference . . . 51
1.6 Conclusion . . . 57

2 Impact of anti-obesity medication initiation and duration on weight loss in a comprehensive weight loss program 60

3 Mobile Engagement Behaviors Associated with Sustained 15Weight Loss 71

Bibliography 79

List of Figures

1.1 Canonical Correlation Analysis (CCA) 19
1.2 GENIE3 GRN inference 22
1.3 sci-GRN inference workflow 24
1.4 PBMC scRNA and scATAC integration 28
1.5 Regulon activity inference. 33
1.6 Stages of cardiac morphogenesis (image borrowed from Basu et al. 2018[1] 38
1.7 Heart sample generation 39
1.8 scATAC clustering using ArchR's iterative-LSI 40
1.9 Human heart scRNA and scATAC integration. 45
1.10 RE1 enhancer in proximity of atrial specific genes, NPPA and NPPB . . 46
1.11 Regulon activity binarizartion 52
1.12 Regulon activity inference. 54
1.13 The regulon activity specificity 56

List of Tables

1.1 Top 10 targets of PAX5 obtained by sci-GRN 34
1.2 GO enrichment analysis using PAX5 ranked targets 35
1.3 PAX5 target genes obtained by only using GENIE3 GRN inference (without providing a base) 36
1.4 Annotated scRNA-seq cell types 43
1.5 Heart GRN base summary 47
1.6 Top 10 targets of TBX5 predicted by sci-GRN 49
1.7 Ranked-base GO enrichment analysis using TBX5 target genes 50
1.8 GO enrichment analysis using a random ranking of TBX5 targets 50

Abstract

Integrative analysis of single-cell RNA and single-cell ATAC profiles to infer gene regulatory networks

by

Rojin Safavi

The human heart is one of the first organs to form during embryogenesis. By week 3, heart is a linear tube; then, between week 3 and 8, it undergoes a complex looping process to form heart chambers. From week 8 up until birth, the heart undergoes a period of rapid growth and expansion as the various cardiovascular lineages mature to support the developing fetus. Defects in this process lead to congenital heart defects. Therefore, understanding heart development at the molecular level will give insight into the mechanisms leading to the disease, and principles can further be applied for regenerative medicine purposes. In addition to genetic factors, lifestyle also significantly contributes to maintaining a healthy heart; therefore, lowering the risk factors that contribute to heart diseases such as obesity is vital. This thesis investigates computational and statistical approaches that help understand the influences of molecular, genetic, and lifestyle components in maintaining a healthy heart.

Single-cell sequencing has provided a unique opportunity to study the heart at cellular resolution. Recent advances in single-cell sequencing allow scientists to sequence different modalities, such as proteomics, spatial transcriptomics, methylation, and chromatin accessibility at the level of individual cells. These advances offer a unique

opportunity to develop computational methods for integrative single-cell data analysis to benefit from the information in multiple modalities. In the first chapter of this thesis, I present Single-cell Integrative Gene Regulatory Network (sci-GRN) inference framework that uses single-cell transcriptomics and epigenomics to infer gene regulatory networks (GRNs), and I demonstrate its application to the human heart development. Here, I first describe how each omic is processed independently to cluster the cells. Next, I describe how we annotated cell types from scRNA-seq data using known marker genes. Third, I present how I used the scRNA-seq data to infer the expression and cell type of each ATAC cell. Lastly, I present how sci-GRN uses the integrated data to build a gene regulatory network.

In addition to genetic factors, lifestyle also significantly contributes to maintaining a healthy heart; therefore, lowering the risk factors contributing to heart diseases is vital. Studies have shown that obesity is associated with an increase in cardiovascular diseases. Therefore, it is crucial to tackling obesity systematically. Enara Health is a medical company based in San Mateo, California that offers a unique hybrid digital and in-person weight loss program. By combining mobile technology with healthcare visits, the program provides patients with personalized physician-driven medical weight loss programs to tackle obesity[2].

In the second chapter of my thesis, I present our study on analyzing one of the cardiovascular diseases" risk factors, obesity. I demonstrate how my statistical analysis implies the significant role of anti-obesity medication initiation and duration on weight loss in obese patients enrolled in Enara's weight loss program.

In the final chapter, I further analyze the impact of Enara's weight-loss program mobile application (mHealth) on weight loss in obese patients. The mobile application under study allows for examining and comparing self-monitoring, feedback, support, and educational features within the same mHealth program, which may shed insight into which features and usage are better for sustained weight loss[2].

To my parents, Azita and Ali,

because of your unconditional love and support,

To my little big sister Kimia!

because I love you so much!

To my husband Omid,

because you have always seen the best in me.

Chapter 1

sci-GRN: Integrative Gene Regulatory Network Inference

1.1 Abstract

Recent advances in single-cell technology have enabled sequencing of chromatin accessibility in addition to transcriptomic sequencing. These advances offer a unique opportunity to develop computational methods for integrative single-cell data analysis. Here, I present Single-cell "Integrative Gene Regulatory Network" (sci-GRN) workflow that uses scATAC-seq and scRNA-seq to infer gene regulatory networks (GRNs). I benchmarked sci-GRN framework's ability to infer known GRNs in peripheral blood mononuclear cells (PBMC), and demonstrated its application to fetal heart data. I further tested the efficacy of sci-GRN in finding known targets of PAX5 in PBMC and TBX5 in fetal heart data. Furthermore, I validated the model's capability in recapit-

ulating a recently discovered cardiac enhancer, RE1, which regulates the expression of NPPA and NPPB in cardiomyocytes. Finally, I utilized the inferred network to investigate the regulon activity and specificity across different cell types using univariate linear models. Together, the results demonstrate the capability of the sci-GRN framework to infer the GRNs through integrative analysis and provide a framework to predict novel regulons across the cell types.

1.2 Introduction

Gene regulatory networks (GRNs) represent the interaction between transcription factors (TFs) and target genes. Constructing a valid GRN is essential for interpreting biological processes at molecular levels[3]. Network-based analysis has been applied to different areas of research[4, 5] such as finding disease-related genes[6, 7, 5] and understanding cell functioning[8, 4]. Building a GRN inference model has been an active area of research for the past 20 years[4]. Network inference models were initially developed for inferring GRNs from bulk transcriptomic[3]. Advances in sequencing technology have now enabled researchers to sequence at the level of individual cells[3, 9, 10]. Unlike bulk data, single-cell sequencing does not obscure cellular heterogeneity by averaging the omic across samples[3, 11]; therefore, it provides a deeper understanding of cellular heterogeneity[4].

The majority of single-cell GRN inference methods infer the GRN from transcriptomic data[11, 12]. Inferring GRN from single-cell RNA expression is challenging,

mainly due to the high drop-out rate and technical noise[13, 14]. Recent benchmarking of some GRN inference models using single-cell RNA data has demonstrated imprecise results, with GENIE3[15] having the better performance and stability overall[16, 11]. One possible explanation is that single-cell RNA per se may not carry sufficient information for constructing a GRN, and it may be crucial to integrate information from other modalities or databases[13, 14]. SCENIC[17] tackles this issue by post-processing GENIE3 weighted-adjacency output using a transcription factor binding site (TFBS) database[13]. Based on the database, only interactions are kept that the TFBS is enriched in the target gene's promoter[13]. Although SCENIC prunes potential false-positive edges, at the same time, it does not account for distal enhancer-dependent regulations. Studies have indicated the importance of enhancers in gene expression regulation[18]. For example, more than half of the enhancers in Drosophila are over 10kb from the gene that it regulates[18, 19].

Gene transcription depends on the ability of transcription factors that bind to short sequences of DNA (5-15 base pairs) called motif[20]. Different transcription factors have different affinities for binding to different motifs. Motifs are abundant in the genome but, depending on the cell type and the chromatin structure around the gene's promoter or the enhancer, they are not always physically available to recruit transcription factors[21, 22]. Therefore, information regarding motif accessibility is crucial for understanding transcription factor activity across different cell types.

Recent advances in single-cell sequencing technology have equipped scientists to measure the chromatin structure at the level of a single cell through transposes-

accessible chromatin sequencing (scATAC-seq)[23, 24]. scATAC-seq provides valuable insights into regulatory elements such as promoters and enhancers in individual cells[12]. Therefore, scATAC-seq data provides another layer of information that could be valuable for making a more informative GRN inference.

Here, I present "Integrative Gene Regulatory Network" (sci-GRN), a novel framework that infers GRNs by integrating scRNA and scATAC data. sci-GRN workflow consists of building an initial gene regulatory network containing a comprehensive set of putative connections between TFs and their targets through scRNA and scATAC integration and assigning importance based on the weighted interactions predicted by the GENIE3 random forest algorithm. I benchmarked the model using PBMC dataset and further demonstrated its application to the human heart datasets. I investigated the inferred GRNs through multiple lenses. First, I demonstrated the model's capability in retrieving a recently discovered enhancer, RE1, in heart data. Second, I used decoupleR univariate linear model[25] to infer regulon activity and specificity across cell types. Next, I tested the capability of the inferred activity matrix in clustering cells. Furthermore, I did a literature review on the top inferred targets of well-studied transcription factors, TBX5 (heart) and PAX5 (PBMC), and performed GO enrichment analysis on the ranked targets of individual TFs to validate the expected biological processes.

1.3 Background

1.3.1 RNA sequencing reveals gene regulatory networks in development and disease

Gene regulatory networks (GRNs) record the interactions between transcription factors and target genes. Seminal studies on sea urchins' developmental GRNs are examples of the profound insights unraveled by GRNs. Sea urchins' developmental GRNs studies demonstrated that the specification and identity of most cells in sea urchins could be explained at the level of GRNs[26, 27, 28, 29]. GRNs define cell identity, fate, and function by controlling gene expression[29]. Disruption in the GRN regulation could play an essential part in human disease[30]. A bulk RNA study conducted by Zeng et al. investigated the contribution of GRNs in heritability of single nucleotide polymorphisms (SNPs) that are associated with expression levels in coronary artery disease[31], and identified 28 GRNs involved in active coronary artery disease[31]. Advances in single-cell sequencing have drawn interest in reconstructing gene regulatory networks using single-cell data. Kathiriya et al. investigated how TBX5 transcription dosage translates into altered GRNs in human iPSC-derived cardiomyocytes from scRNA-seq data[32]. By conducting a time-point matched differential expression analysis, Kathiriya et al. compared the inferred GRN of control to $TBX5^{in/+}$ or $TBX5^{in/del}$, and identified nodes that had TBX5-dependent reduction in pagerank centrality. They identified TBX5 and MEF2C interaction through their analysis and further validated the interaction in mice[32]. Therefore, GRNs play critical roles in maintaining cell-type

states and function by dictating what genes and in what amount should be expressed in different cell types[33].

A gene is expressed when a transcription factor binds to its promoter or distant regulatory elements such as enhancers. Promoters are around the gene's transcription start site (TSS), while enhancers could be as far as 1 Mbp from the gene's TSS[34]. The binding of a transcription factor to the gene's regulatory element leads to RNA polymerase II recruitment at the gene's TSS and transcription initiation. The tight correlation of a TFs expression with another gene's expression could suggest the TF regulates the gene. For example, GENIE3 GRN inference predicts the expression profile of each target gene using the expression of transcription factors[35]. However, establishing a direct relationship between a TF and another gene is a major challenge in GRN inference as strong correlations can be present among those indirectly regulated genes (e.g., the factor regulates a second factor that then regulates the gene)[13, 14]. SCENIC tackles this issue by post-processing GENIE3's output using a transcription factor enrichment database [17]. Based on this dataset, SCENIC retains interactions with the TFBS enriched in the target gene's promoter. Although SCENIC prunes potential indirect targets, it does not consider distal dependent regulations. Determining which TF binding sites are utilized in any condition remains an open problem. What is clear is that the accessibility of these motifs is one of the keys to distinguishing the truly functional sites occupied by TFs in different cell types.

Transcription factors have DNA binding domains that could recognize specific regions of the genome called motifs. Motifs are short sequences of DNA (5-15

base pairs)[20] abundant across the genome. Transcription factors bind to motifs that are similar to one another but not necessarily identical[36]. Motifs are not always physically accessible to transcription factors because of the DNA arrangement. DNA is wrapped around specific protein complexes called histones. Depending on the histone post-translational modifications, histones can relocate and make certain parts of the genome more or less accessible[37]. For example, bisulfite sequencing of histones has shown that active/accessible promoters have high H3K9ac modification, and active/accessible enhancers have high H3K4me1/H3K27ac modification[38, 33]. Thus, a motif could be accessible or inaccessible depending on the chromatin states. DNA accessibility can be measured using the assay for "transposase-accessible chromatin sequencing" (ATAC-seq)[24]. ATAC-seq isolates open chromatin regions by treating the DNA with hyperactive Tn5 transposase[24]. Thus, ATAC-seq allows researchers to identify potential regulatory regions of the genome. Together, scRNA-seq and scATAC-seq provide complementary information for building a more robust GRN.

1.3.2 Discovering transcription factor motifs reveals regulatory linkages

Motif discovery is the problem of finding short sequences of DNA that have similar biological function[36]. There are over 100 motif discovery algorithms built over the past decades[39]. MEME (Multiple EM for Motif Elicitation)[40] is one of the most popular tools used for motif discovery. MEME algorithm is based on optimizing a Position Weighted Matrix (PWM). PWM is a matrix of the size 4Xm where m is the

length of the motif, and rows are A, T, C, and G. Each position in the PWM matrix represents the probability of observing the nucleotide in that position[39]. The idea behind MEME algorithm is to optimize the PWM matrix by Expectation-Maximization (EM) to discover potential motifs[39].

Motif discovery per se is not an indication that a transcription factor binds a motif. Depending on the histone post-translational modifications, histones can relocate and make certain parts of the genome more or less accessible[37]; Thus, a motif could be accessible or inaccessible depending on the chromatin states. One approach to learning about these open regions is chromatin immunoprecipitation sequencing (ChIP-seq). ChIP-seq allows researchers to identify accessible genomic regions occupied by specific transcription factors[39]. ChIP-seq assay requires designing antibodies with high sensitivity and specificity for the TF of interest[41]. Once the antibody is available, it needs to be tested in the cell line under the study to obtain TF-bound regions. Human genome is estimated to encode around 1,639 TFs[42]. One goal of consortiums such as ENCODE[43] and Roadmap Epigenomics[44] is to build a comprehensive database of regulatory elements in all cell-type, but such data is not available for all TFs and cell lines to date (339 transcription factors have been characterized in 139 cell lines in ENCODE project[43]). Thus, new systematic approaches are needed to probe TF binding in cell types of interest.

ATAC-seq promises to complement these approaches since only the binding motif of the factor needs to be characterized rather than having an antibody to probe where the factor is binding. ATAC-seq is state of the art for identifying accessible

chromatin regions in the whole genome[24]. Unlike ChIP-seq, ATAC-seq can capture accessible regions of the genome in one shot, and it is not transcription factor specific. Thus, designing an ATAC-seq experiment is much cheaper, and it could provide valuable information for studying regulatory regions of the genome.

1.3.3 Single-cell sequencing

The majority of transcriptomic and epigenomics analyses are based on the assumption that cells in a tissue are homogeneous[45]. Such analyses are more likely to miss cell-to-cell variability[45]. Recent advances in DNA sequencing technology have enabled researchers to study tissues at the level of cellular resolution. Single-cell RNA sequencing (scRNA-seq) is state of the art for analyzing single-cell transcriptomic, and single-cell ATAC sequencing (scATAC-seq) is state of the art for analyzing the regulatory landscape in single-cells[46].

1.3.3.1 Single-cell RNA sequencing implicates heart development gene regulatory networks

The sequencing of the entire transcriptomics at the level of individual cells is called single-cell RNA sequencing (scRNA-seq)[45]. scRNA-seq data opens up new avenues for an unbiased study of cellular heterogeneity within a tissue[45, 4, 47]. This type of data helps researchers investigate different biological questions such as discovering rare cell populations in a tissue[45], identifying novel marker genes[47], switches in cellular states, and inferring regulatory networks[47]. For example, Kathiriya et al.

examined how TBX5 haploinsufficiency affects GRNs in human iPSC-derived cardiomyocytes scRNA[32] and identified TBX5-dependent interactions critical for congenital heart diseases[32]. Therefore, scRNA-seq provides valuable insight into understanding heart development through the lens of gene regulatory networks.

1.3.3.2 Single-cell ATAC sequencing reveals heart reprogramming motifs

The single-cell scATAC-seq allows identifying open (accessible) chromatin regions at individual cell resolution[48]. Similar to bulk ATAC-seq, scATAC-seq uses Tn5-transposase to identify open regions of the chromosome[49]. Tn5-transposase preferentially binds to accessible regions of the genome and simultaneously tags and fragments the DNA[49]. scATAC-seq enables scientists to study cell-to-cell chromatin accessibility variation[48], find rare cellular subpopulations[49], and predicting potential candidate enhancers and regulatory elements[50]. scATAC-seq conducted by Wang et al. on the early stage of induced cardiac reprogramming (iCM) cells identified a bimodal role of SMAD3 transcription factor at different stages of cardiac reprogramming through studying the motif accessibility[51]. Therefore, scATAC-seq provides valuable insights into identifying accessible genomic regions across developmental trajectories and cell types.

1.3.4 Gene regulatory inference

Gene regulatory networks consist of transcription factors and their target genes. Gene regulatory network inference refers to identifying such interaction through analyzing genomic data. Network-based analyses have been applied to different ar-

eas of research[4, 5] such as finding disease-related genes[6, 7, 5], understanding cell functioning[8, 4], etc. GRN-based analysis help scientist answer a variety of biological and medical questions[52]. For example, GRN inference could assist researchers in building novel hypotheses about TF-target interactions. There are over 20,000 protein-coding genes and around 1,600 transcription factors in humans. GRN inference narrows down the number of interactions to the ones that are more significant[52]. Such knowledge helps researchers in designing proper wet-lab experiments to test the generated hypothesis (such as designing a ChIP-seq experiment)[52]. Furthermore, GRNs could be used as guides for perturbation experiment design[52]. For example, based on the GRN inference, specific perturbations could experimentally be introduced to boost a specific signal (such as pathway) or activity (such as regulon)[52].

Developing GRN inference models have been an active area of research for the past 20 years[4]. One of the popular branches of GRN inference is regression-based models[12]. Regression-based inferences are based on the idea that expression of a target gene could be predicted by the expression of transcription factors regulating it[12]. GENIE3 is a popular regression-based inference algorithm, which was shown to have the best performance in DREAM4 in-silico network challenge[35, 53]. GENIE3 assumes that the expression of each gene is a function of the expression of some transcription factors[12]. GENIE3 attempts to solve this function for each gene using Random Forest regression. Furthermore, Random Forest regression provides an importance measure for each TF-gene interaction, which indicates how important the connection between the two genes is.

GRN inference from scRNA-seq is challenging due to the high drop-out rate and technical noise[13, 14]. Recent benchmarking of some GRN inference models using single-cell RNA data has demonstrated imprecise results, with GENIE3[15] having the better performance and stability overall[16, 11]. The low performance could potentially be the consequence of using scRNA-seq alone for the GRN inference[11]. For example, in the case of regression-based inference models, a correlation between a transcription factor and a gene does not necessarily mean that the gene is a target of the transcription factor; thus, using single-cell RNA alone could lead to false-positive edges. SCENIC tackles this issue by post-processing GENIE3's output (adjacency matrix) using a transcription factor enrichment database[13]. Based on this dataset, SCENIC only keeps interactions that the TFBS is enriched around the promoter of the gene that it is correlated to[13]. Although SCENIC prunes potential false-positive interactions through post-processing, at the same time, it is limited to the area around the promoter; therefore, it does not take into account distal regulatory elements such as distal enhancers.

Cicero is the front-runner method for identifying cis-regulatory interactions (such as enhancer and promoter interaction) using scATAC-seq data[54, 55]. Cicero calculates peak-to-peak correlation by first creating groups of highly similar cells through sampling. Then, it computes the covariance between each pair and estimates the correlation using graphical LASSO, penalizing distant pairs more than proximal ones[55]. Through Cicero analysis, one can find peaks that are correlated to peaks in a gene's promoter and build a regulatory map. However, it should be noted that an accessi-

ble promoter does not necessarily mean that the gene is being expressed[50]; therefore, scRNA-seq could provide complementary information to such analysis.

GRN inference using multi-omic data, such as bulk RNA-seq and ATAC-seq, has been demonstrated to enhance the GRN inference accuracy significantly[56, 57, 58, 13]. GRN inference from multi-omic single-cell data is still in its infancy[13]. The majority of single-cell GRN inference methods (such as ScoMAP[50]) require that the two omics data types are applied to the same cell[12]. Recent innovations in single-cell sequencing allow simultaneous profiling of transcriptomics and epigenomics for each cell (such as sci-CAR[58])[13]. However, most publicly available single-cell data are not from joint sequencing. Some methods such as Canonical Correlation Analysis (CCA)[59, 13] provide a framework to predict the RNA expression for each ATAC cell. The inferred RNA expression, along with the ATAC information, could then be used for GRN inference[13].

Seurat software package has provided a CCA implementation to integrate scRNA and scATAC data[59]. Seurat's CCA integration approach has been widely used for scRNA and scATAC integration [60, 59, 61, 62, 63]. Canonical Correlation Analysis assumes a linear correlation between scATAC and scRNA cells. Furthermore, it presumes that the cells from the two modalities have similar measurements (for example, having genes as rows). Therefore, the scATAC peak level must be converted to gene level to integrate the two modalities. Different methods have been introduced to convert the scATAC peak level to a gene score level[62, 59]. ArchR's benchmarking of 50 different gene scores calculations models demonstrated a consistent better perfor-

mance of models that include the whole gene body accessibility, exponential weighting function, and gene boundaries to measure the gene activity scores[62]. Therefore, in addition to linearity and having similar row names, CCA also assumes that the gene activity scores represent the chromatin state structure.

In recent years other approaches have been introduced for scATAC and scRNA integration. SOMatic uses self-organizing maps (SOM), a type of neural network, to integrate scRNA and scATAC data. In brief, SOMatic builds SOM for each modality and combines them using a linking function[64]. SOM model convergence might not be reached untill hundreds of thousands of iterations; therefore, it could computationally be expensive[12]

In addition to developing computational methods for GRN inference, consortiums like Roadmap Epigenomics[44] and ENCODE[65] attempt to characterize regulatory elements of the genome through wet-lab experiments. Roadmap Epigenomics aim is to provide reference epigenomic data from human tissues and cell lines, such as DNA methylation and ChIP-seq[44]. ENCODE consortium also produces epigenomic data such as methylation, ChIP-seq, and ATAC-seq of human and mouse, in addition to integrating data from other resources into their database[65].

GRN inference in the heart is an active area of research. Hocker at al. used the Cicero peak coaccessibility method to find cell type-specific candidate cis-regulatory elements (cCRE) in human heart[66]. They further examined the existence of cardiovascular disease-associated variants in the cCREs by conducting a GWAS variant enrichment analysis. They reported a significant enrichment of atrial fibrillation variants

in atrial and ventricular cCREs[66]. In another study, Kathiriya et al. examined how TBX5 dosage influences the predicted GRNs in human iPSC-derived cardiomyocytes scRNA[32]. They measured changes in node pagerank centrality across the developmental time points by comparing the GRN of control to $TBX5^{in/+}$ or $TBX5^{in/del}$, and identified a significant pagerank centrality reduction in MEF2C, a TBX5 target gene. They further validated the predicted TBX5-MEF2C interaction in mice[32].

1.4 Methods

1.4.1 Heart tissue samples

Our collaborators in the Chi lab at University of California, San Diego., collected de-identified heart tissue samples with previous patient consent in strict observance of the legal and institutional ethical regulations. Protocols were approved by the #101021X (Institutional Review Board) at the University of California, San Diego.

1.4.2 scRNA-seq processing

The RNA-seq processing was performed by our collaborators in the Chi lab at the University of California, San Diego. Cell Ranger v3.0.1 was used to process raw sequencing data. In brief, Cell Ranger generated a count matrix for each tissue sample by demultiplexed reads and aligning them to the human hg38 reference genome. Next, the count matrices were merged and processed using the Seurat v4.0.1 R package[67]. Seurat is a comprehensive R package for analyzing scRNA data. Seurat

tutorials (https://satijalab.org/seurat/) were followed for further downstream analysis. Low-quality cells (with less than 1000 genes detected) and dead cells (with mitochondrial read percentage of $> 30\%$) were removed from downstream analysis. Potential doublets were detected (with a doublet rate of 5%, recommended by 10X Genomics) and removed using DoubletFinder (https://github.com/chris- mcginnis-ucsf/DoubletFinder). Gene expression was normalized by dividing the gene count for each cell by the total count for that cell and taking a natural log transform using the NormalizeData function. The top variable genes were detected with the FindVariableFeatures function with default parameters. Next, a linear transformation was used to center the data (mean of 0 and variance of 1 across the cells) using the ScaleData function. Principal components were computed using the RunPCA function and were used for generating a nearest neighbor graph. The graph was then used to cluster cells using the FindClusters function. Cells were projected into two dimensions using the runUMAP function. Cluster-specific markers were identified by the FindAllMarkers function, and cell identities were assigned to each cluster based on known cell type markers.

1.4.3 scATAC-seq processing

Raw sequencing reads were processed using Cell Ranger ATAC v2.1.0 pipeline from 10X Genomics. Cell Ranger ATAC outputs a bed file called a fragment file. A fragment file stores the coordinates of the aligned reads in each cell. After generating the fragment files, ArchR v1.0.2 was used to process the fragment files. ArchR[60] is a comprehensive R package for analysing single-cell ATAC data. ArchR's createArrow-

Files with the default parameters was used to build the ArchR object, as described in the tutorial (https://www.archrproject.com/bookdown/). Dead and low-quality cells were eliminated by the number of fragments (less than 1000) and low TSS enrichment (less than 4). TSS enrichment score is calculated by dividing the read coverage at the center of TSS by the coverage around the TSS (+/-2kbp from the TSS)[60]. Cells were clustered using ArchR's addIterativeLSI function with the default parameters, then projected into two dimensions using addUMAP with the default parameters. For each ATAC cell, the cell type and expression were inferred using addGeneIntegrationMatrix function (see Methods, Section scRNA and scATAC integration section). Peak-calling was performed using MACS2 (with p-value=0.001)[68] to obtain peak enrichment across each cell. Finally, addPeak2GeneLinks function was used to measure the correlations between peak accessibility and gene expression (see Methods, Section gene regulatory network base section).

1.4.4 scRNA and scATAC integration

The sci-GRN workflow consists of 2 main parts, building the GRN base where transcription factors are connected to their putative target genes (see Methods, Section Gene regulatory network base section) and assigning importance to each interaction (see Methods, Section Putative target gene ranking section). In sci-GRN, building the GRN base is dependent on having a joint dataset in which each cell has ATAC and RNA measurements. Although new approaches produce a joint dataset, many experiments collect scRNA-seq independently than scATAC-seq, and thus an extra step is needed to

connect the two datasets. sci-GRN uses Seurat's[59] integration approach, implemented in ArchR[60] to infer the expression and cell type of each ATAC cell. Seurat integration method is based on Canonical Correlation Analysis (CCA)[59]. CCA projects cells from the two modalities into another space, where in that space, the two modalities are highly linearly correlated to one another[69]. The prerequisite of CCA analysis is having the same row names (in this case, genes). ArchR converts the accessibility data into gene scores to match the rows (Figure 1.1). ArchR builds the Gene Score Matrix by identifying a subset of peaks that overlap a gene, then converts the distance between each peak and gene body to a distant weight e(-abs(distance)/5000) + e-1, and finally adjusts for large genes by multiplying the distant weight by (1/ gene size). Finally, the scRNA and the Gene Score Matrix are used as CCA inputs to find the closest scRNA cell to each ATAC cell (Figure 1.1). The inferred label and expression are obtained for each ATAC cell based on the k-nearest neighbors in this new dimension (Figure 1.1).

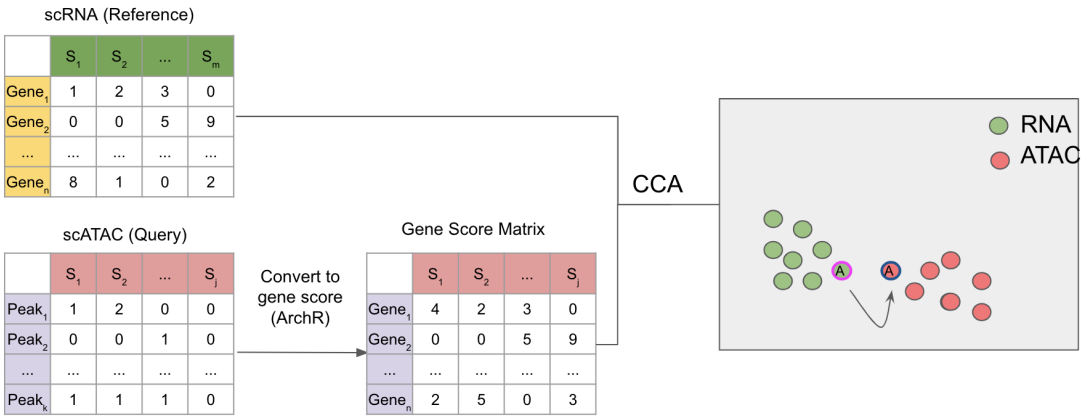


Figure 1.1: Canonical Correlation Analysis (CCA)

scRNA is The reference count matrix with "n" genes along the row and "m" samples along the column. The scATAC matrix is the query matrix with "k" peaks in the rows and "j" samples in the column. The scATAC peak matrix is converted to a Gene Score Matrix with "n" genes along the row and "j" samples along the column using ArchR. The scRNA and Gene Score Matrices are given as inputs to CCA. The gray box represents the mapped cells from the two modalities into a highly correlated space. In the new space, green dots are scRNA cells, and red dots are ATAC cells. Based on the proximity of ATAC cells and scRNA cells in the new space, scATAC cell type and expression are inferred (for example, the ATAC cell with a blue outer edge is in the proximity of the RNA cell "A" with a purple outer edge; therefore, it probably has the same cell type and expression as RNA cell "A").

1.4.5 Gene regulatory network base

The next step in sci-GRN is to infer an initial gene regulatory network containing a comprehensive set of putative connections between TFs and their targets, henceforth referred to as the GRN base. sci-GRN reconstructs a GRN with the as-

sumption that TFs interact with their target genes only by binding to sites located in regions of the genome detected as accessible[70]. sci-GRN leverages the joint dataset reconstructed in the previous step to compute cis-interaction correlations between the inferred expression of a gene and the accessibility of cis-regulatory elements (using ArchR's addPeak2GeneLinks function). The cis-interaction correlation could potentially connect regulatory elements to their putative target genes. To narrow down the search space and focus on the known biology represented in the GRN, only peaks with known TF binding sites (using CisBP database[19])) are retained. Next, relevant cell-type-specific sites in the genome are retained as those demonstrating a high correlation (default ArchR's threshold, Pearson >0.45) and low FDR value (< 0.001) (Figure 1.3b). In addition to cis-interaction correlation, sci-GRN computes the trans-interaction correlation between TF-inferred expression and their accessibility in cis-regulatory elements (Figure 1.3b). Trans-interaction correlation is computed by creating a feature $\times$ cell-type matrix for each omic and computing the Pearson correlation between the two feature vectors (Figure 1.3b). Finally, in order to retain only peaks with at least one associated TF, sci-GRN retains peaks with high trans-interaction correlation (Pearson >0.6) (Figure 1.3b). The final table is a binary table with non-zero entries representing the links between TFs and targets (Figure 1.3c, middle). As shown in Figure 1.3c, this binary table represents putative TF-target interactions and represents the GRN base (Figure 1.3c, left).

1.4.6 Putative target gene ranking

To identify TF-target interactions supported by both ATAC-seq and RNA-seq information, sci-GRN ranks the putative target genes of each transcription factor using the GENIE3 random forest regression model (Figure 1.3d). Based on the genes in the GRN base, a matrix of TF expression is used as the input, and a matrix of all putative target genes is used as the output of GENIE3. For each gene in the GRN base, GENIE3 tries to predict its expression using the expression of a set of transcription factors (Figure 1.2). GENIE3 outputs a weighted adjacency matrix with an importance assigned to each interaction. GENIE3 importance is similar to ranking, with higher importance ranking at the top. To remove less significant interactions, the top 10 regulators of each target gene were picked using the weighted adjacency matrix.

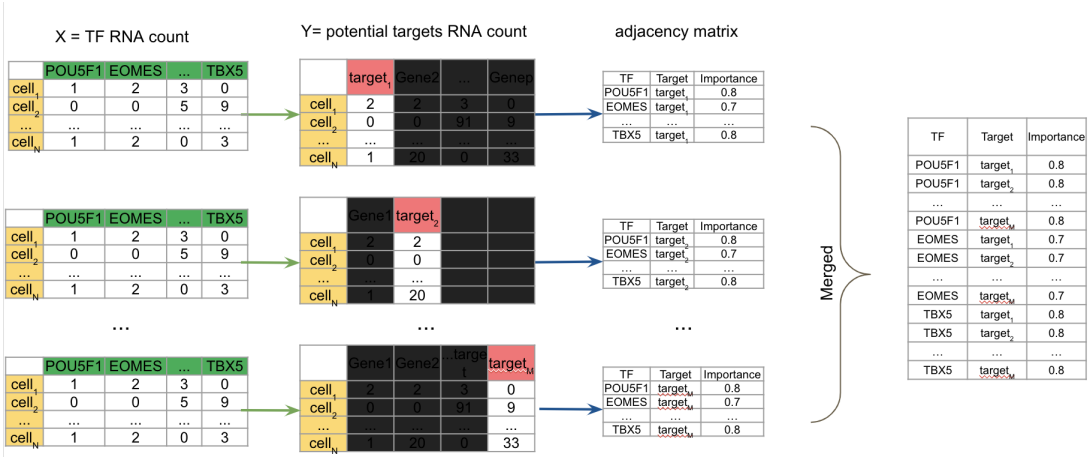
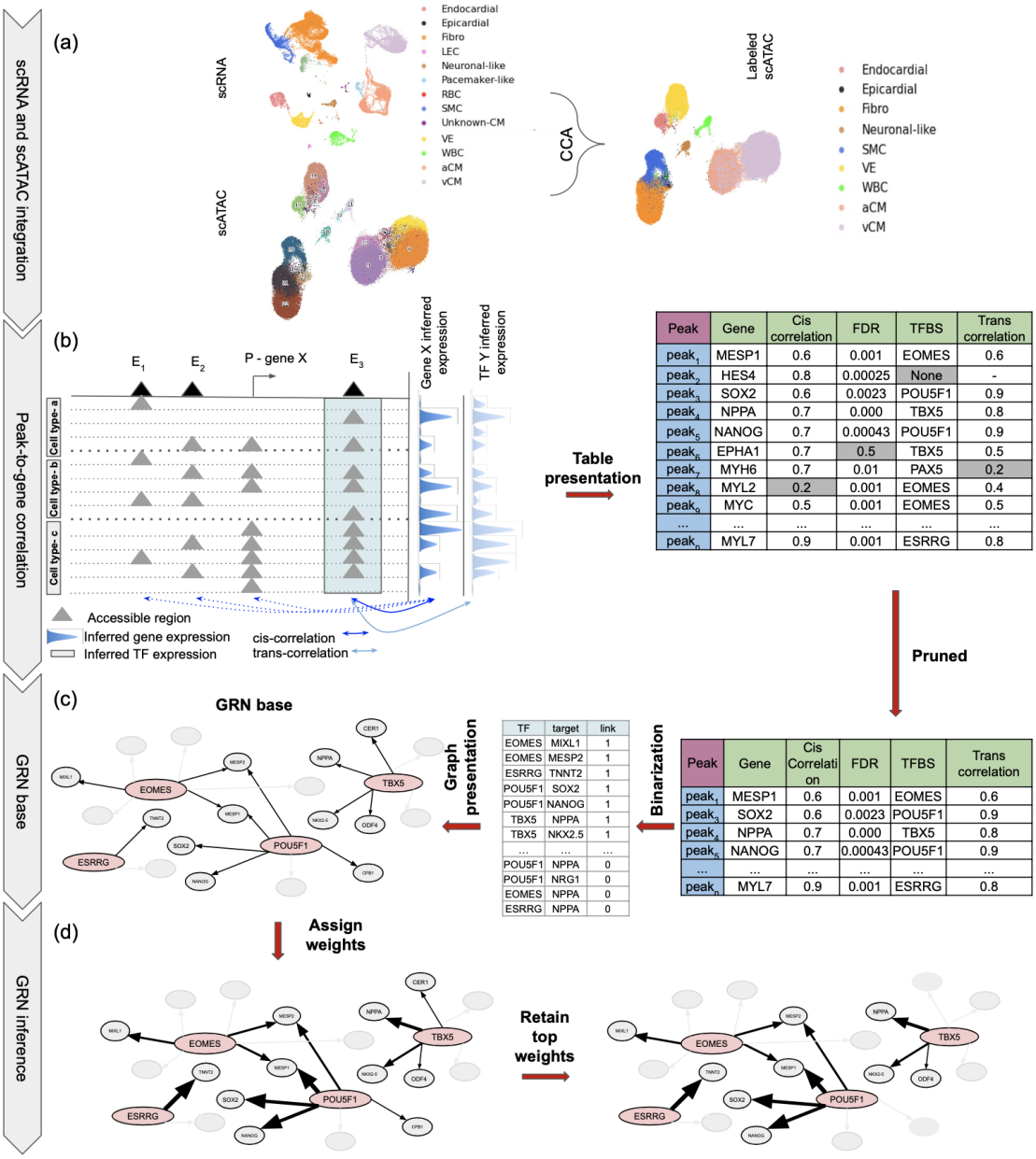


Figure 1.2: GENIE3 GRN inference

The X matrix is an RNA count matrix with cells along the rows and TFs along the columns. Each Y matrix is an RNA count matrix of the putative target gene (predicted by the GRN base) with cells along the rows. For each target gene (M targets total), GENIE3 runs a random forest regression to predict the target gene's expression using the TF expression (green arrows) and outputs (blue arrows) an adjacency matrix with the TFs in the TF column, the target gene in the target column, and the feature importance in the importance column. The final adjacency matrix is built by merging all the adjacencies.



Peak	Gene	Cis correlation	FDR	TFBS	Trans correlation
$peak_1$	MESP1	0.6	0.001	EOMES	0.6
$peak_2$	HES4	0.8	0.00025	None	-
$peak_3$	SOX2	0.6	0.0023	POU5F1	0.9
$peak_4$	NPPA	0.7	0.000	TBX5	0.8
$peak_5$	NANOG	0.7	0.00043	POU5F1	0.9
$peak_6$	EPHA1	0.7	0.5	TBX5	0.5
$peak_7$	MYH6	0.7	0.01	PAX5	0.2
$peak_8$	MYL2	0.2	0.001	EOMES	0.4
$peak_9$	MYC	0.5	0.001	EOMES	0.5
...	...	...	...	...	...
$peak_n$	MYL7	0.9	0.001	ESRRG	0.8

TF	target	link
EOMES	MIXL1	1
EOMES	MESP2	1
ESRRG	TNNT2	1
POU5F1	SOX2	1
POU5F1	NANOG	1
TBX5	NPPA	1
TBX5	NKX2.5	1
...	...	...
POU5F1	NPPA	0
POU5F1	NRG1	0
EOMES	NPPA	0
ESRRG	NPPA	0

Peak	Gene	Cis Correlation	FDR	TFBS	Trans correlation
$peak_1$	MESP1	0.6	0.001	EOMES	0.6
$peak_3$	SOX2	0.6	0.0023	POU5F1	0.9
$peak_4$	NPPA	0.7	0.000	TBX5	0.8
$peak_5$	NANOG	0.7	0.00043	POU5F1	0.9
...	...	...	...	...	...
$peak_n$	MYL7	0.9	0.001	ESRRG	0.8

Figure 1.3: sci-GRN inference workflow

(a) sci-GRN infers the cell type and the expression of each ATAC cell using Seurat CCA analysis. (top, left) UMAP projection of scRNA cells, with cluster names described in the legend. (bottom, left). UMAP projection of scATAC cells with 24 clusters identified by iterative-LSI. (right) Annotated scATAC cells generated from scRNA and scATAC integration. (b) (left) Illustrates accessibility of three putative enhancers (E1, E2, E3) and a promoter (P) in the proximity of gene X. The inferred expression of gene X obtained from scRNA and scATAC integration is plotted as dark blue peaks. The average inferred expression of TF Y averaged across all cells of a cell type (groups of rows) is shown as light blue. Cis-correlation is computed between the accessibility of each cis-regulatory element (enhancer or promoter) with the inferred expression (dark blue double arrows below plot). Trans-correlation is computed between the accessibility of each cis-regulatory element (enhancer or promoter) containing a TFBS with the inferred expression of the TF (light blue double arrows below plot). (right) Table presentation of the interactions. Each Row represents a TF-to-gene interaction. Gray entries do not pass the filtering criteria. (c) (right) Filtered interaction table with each Row representing a putative TF-target interaction. (middle) Binarized TF-gene interactions with 1 indicating a TF-target interaction and 0 indicative of no interaction. (left) GRN base illustration with pink nodes representing TFs and gray nodes with outer black edge representing potential targets. Indirect and redundant targets are shown with an outer edge gray. (d) (left) GRN network with weighted interactions obtained by GENIE3. The edge thickness represents that interaction weight, with thicker arrows bearing more weight. (right) The top 10 regulators (TFs) of each target gene are retained.

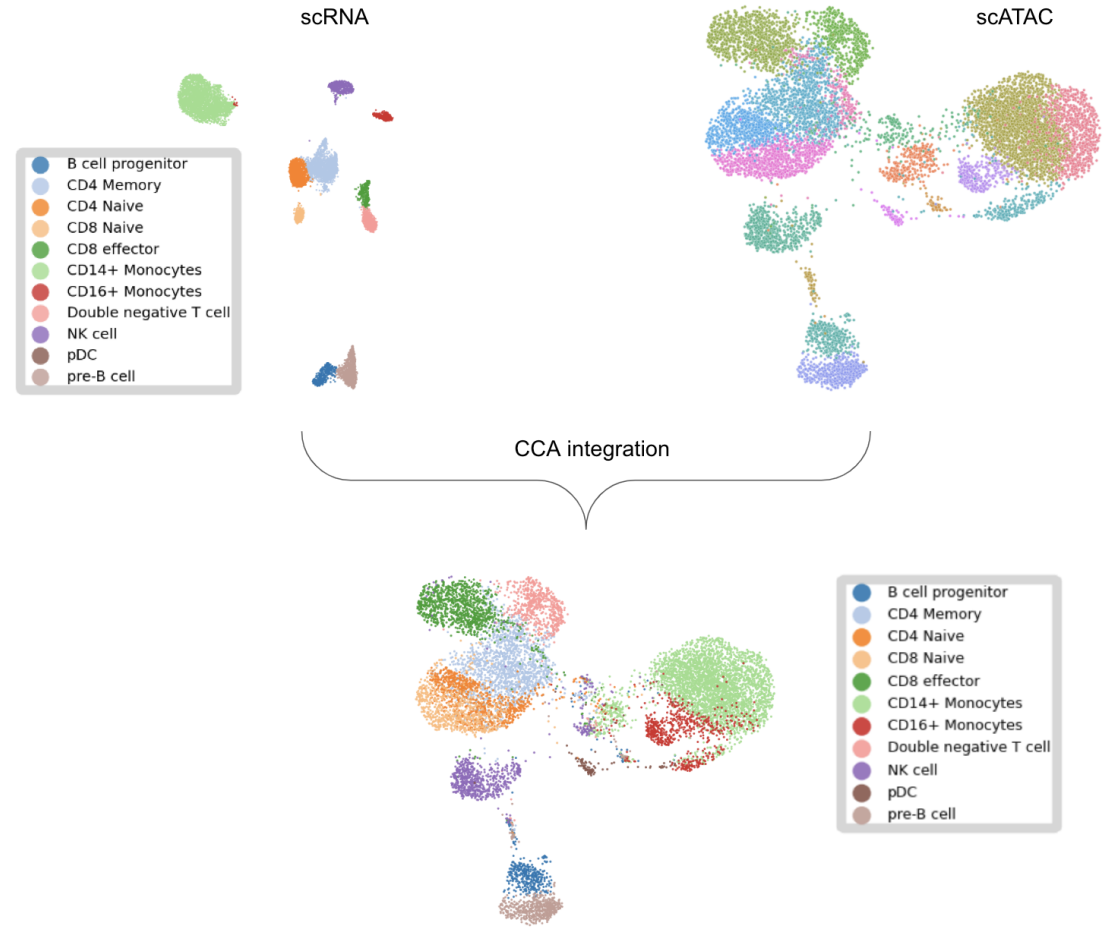
1.4.7 PBMC benchmarking

1.4.7.1 single-cell RNA and single-cell ATAC integration

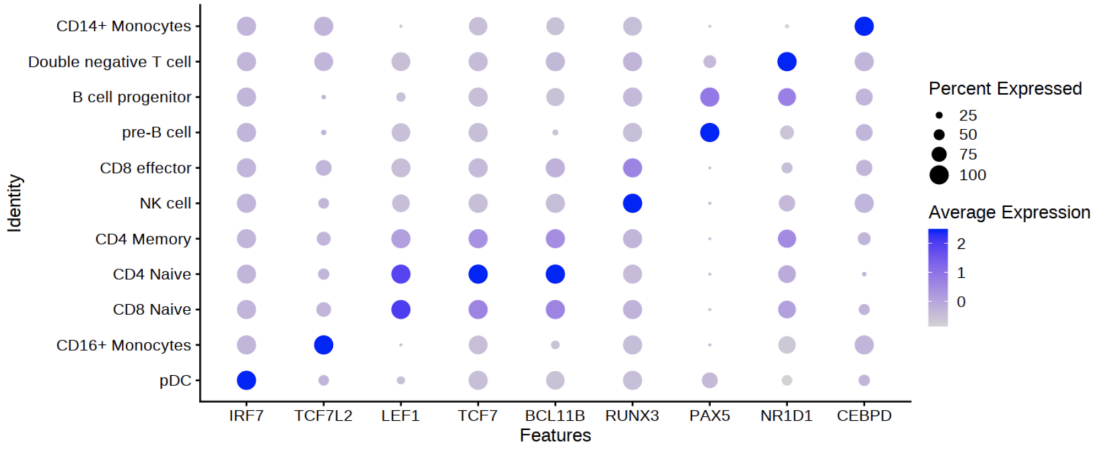
I benchmarked sci-GRN inference framework using a 10X Genomics PBMC dataset[71]. The raw PBMC scRNA data could be found at https://support.10xgenomics .com/single-cell-gene-expression/datasets/3.0.0/pbmc_10k _v3 , and the prepossessed version which was used for the benchmarking could be downloaded from Seurat[59] website (https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.0.0/pbm c_10k_v3). The PBMC scATAC data was downloaded from 10X Genomics site [https://w ww.10xgenomics.com/resources/datasets/10k-human-pbmcs-atac-v1-1-chromium-x-1-1-standard], and processed using ArchR[60] package (see Methods, Section scATAC-seq processing).

Figure 1.4a demonstrates the identified cell-types in scRNA PBMC dataset. PBMC scRNA and scATAC were integrated using Seurat[59] canonical correlation analysis (CCA) (see Methods, Section scRNA and scATAC integration)(Figure 1.4a). The integration allows inferring the inferred cell type and expression of each ATAC cell. The inferred expression and inferred cell type of each ATAC cell were validated using well-known marker genes gathered from the literature. Figure 1.4b demonstrates the inferred expression of cell type-specific markers, and Figure 1.4c represents the expression of the same marker genes in the scRNA-seq data. The dot plot comparison between the inferred expression and scRNA demonstrates similar expression patterns (Figure 1.4c). These results validate the robustness of CCA analysis in inferring cell-type-specific ex-

pression. Therefore, in the absence of a joint dataset, CCA could be used to infer the expression and cell type of each ATAC cell. The resulting inference, along with the scATAC data, could then be used in the sci-GRN inference framework.


scRNA
scATAC
B cell progenitor
CD4 Memory
CD4 Naive
CD8 Naive
CD8 effector
CD14+ Monocytes
CD16+ Monocytes
Double negative T cell
NK cell
pDC
pre-B cell
CCA integration

(a)


CD14+ Monocytes
Double negative T cell
B cell progenitor
pre-B cell
CD8 effector
NK cell
CD4 Memory
CD4 Naive
CD8 Naive
CD16+ Monocytes
pDC
Identity
IRF7
TCF7L2
LEF1
TCF7
BCL11B
RUNX3
PAX5
NR1D1
CEBPD
Features
Percent Expressed
25
50
75
100
Average Expression
2
1
0

(b)

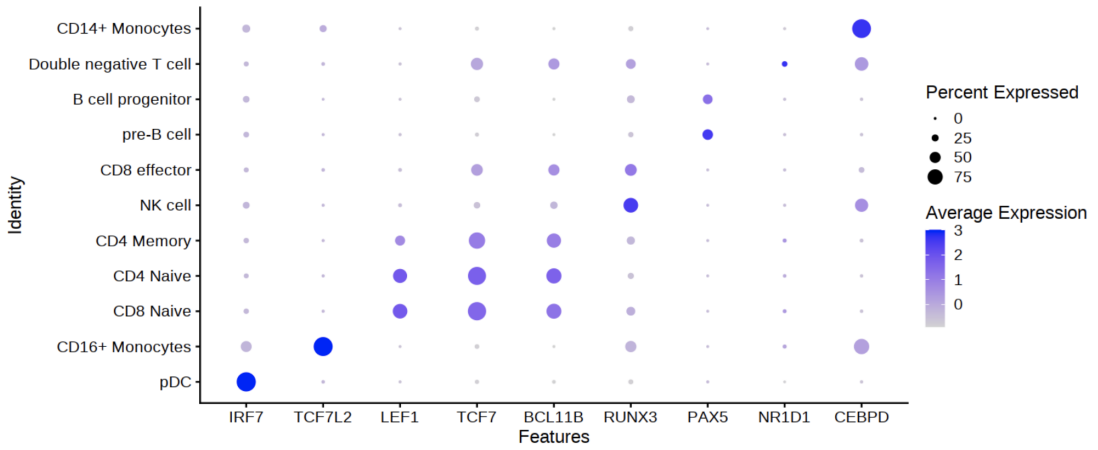


Figure 1.4: PBMC scRNA and scATAC integration

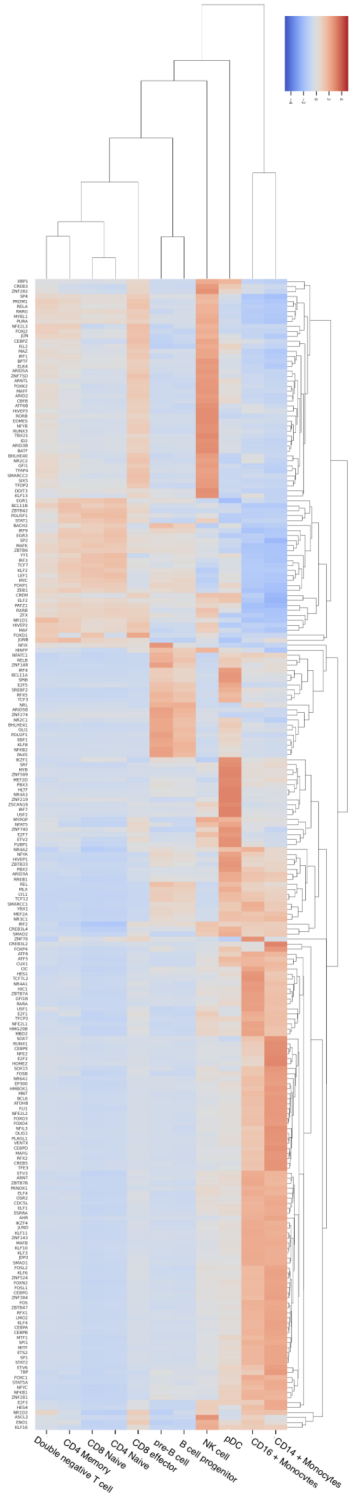
(a) Inferring cell types and expression of each PBMC ATAC cell through scRNA and scATAC integration[59]. The left-top UMAP projection represents annotated scRNA-seq data with each cell type defined in the legend. The left bottom UMAP projection represents the scATAC-seq data clustering obtained by the iterative-LSI. The right UMAP represents the annotated scATAC-seq data. (b) Navigating the expression of marker genes across scRNA cell types using Dot plot. Y-axis represents the annotated clusters, and X-axis represents the marker genes. The size of each dot represents the percentage of cells within the cluster expressing the marker gene. The color encodes the average expression of each marker gene across the cluster. (c) Navigating the inferred expression of marker genes across scATAC cell types using Dot plot. Y-axis represents the annotated clusters, and X-axis represents the marker genes. The size of each dot represents the percentage of cells within the cluster expressing the marker gene. The color encodes the average expression of each marker gene across the cluster.

1.4.7.2 Regulon activity inference

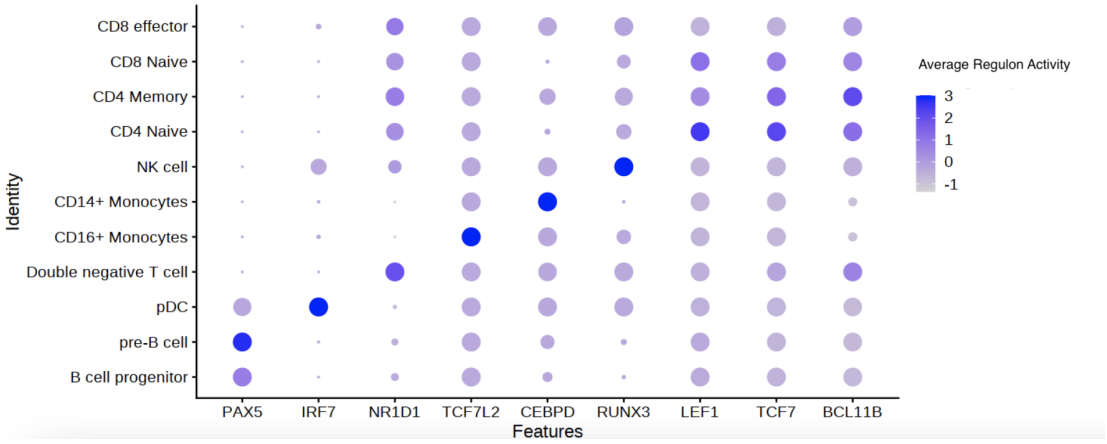
I utilized univariate linear model (ULM)[25] to infer regulon activity from the PBMC inferred GRN. Figure 1.5a represents the mean activity of detected regulons across different cell types. Figure 1.5b shows the activity of some well-characterized regulons across PBMC cells. PAX5 is a master regulator of B cell lineage and plays essential roles throughout B cell development[**?**, 72]. Figure 1.5b demonstrates a high activity of PAX5 in the B cell lineage. IRF7 plays a critical role in pDC cells by facilitating type I interferon (IFN-I) induction[73]. Figure 1.5b indicates a high IRF7 regulon activity in pDC cells which corresponds well with the published literature[73, 74]. Yan et al. reported a high expression of NR1D1[75] transcription factor in double-negative t-cells, which corroborates with our finding that NR1D1 has high regulon activity in double-negative T cells (Figure 1.5b). Wacleche et al. study on CD16+ monocytes cells revealed a unique expression of TCF7L2 (also known as TCF4) in CD16+ monocytes, which corroborates the TCF7L2 inferred expression (Figure 1.5b). CEBPD is a member of CCAAT/enhancer-binding protein (C/EBP) family[76], and plays essential roles in monocyte development and differentiation[76, 77]. Figure 1.5b demonstrates a high activity of CEBPD regulon in CD14+ monocytes. RUNX3 has been reported to have a high expression in NK cells and essential roles in NK cell lineage development[78]. Figure 1.5b demonstrates a high activity of RUNX3 transcription factor in NK cells, which is in agreement with previous studies[78, 79]. LEF1 and TCF7 are two essential transcription factors for naive T cell development. Our result

also indicates a high activity for LEF1 and TCF7 in CD8 naive and CD4 naive cells (Figure1.5b). BCL11B is a key regulator of T cells. A correlation analysis between BCL11A expression and T cell subtypes has shown a positive correlation in naive and memory CD4+ cells[80]. The inferred expression also indicates a higher activity of BCL11A in naive and memory CD4+ cells. Altogether, these findings suggest that sci-GRN detects relevant transcription factor targets; as a result, it can make relevant regulon activity inferences.

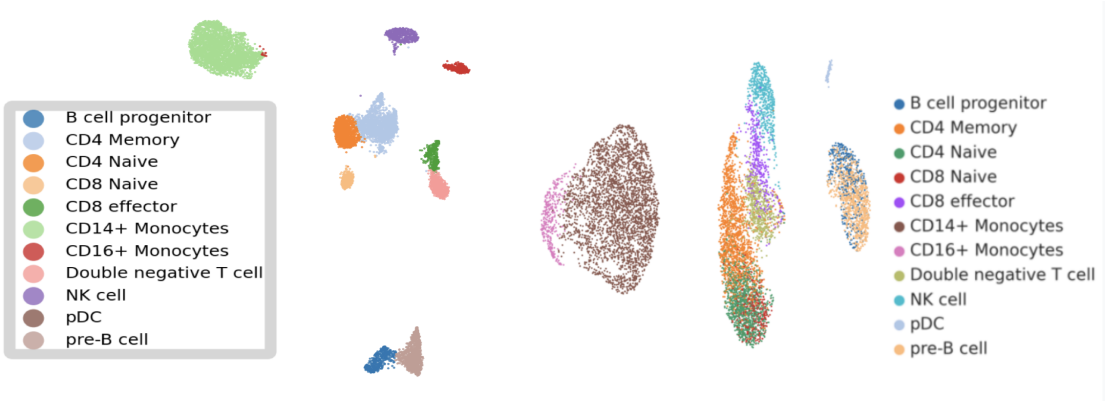
The regulon activity matrix could itself be used for downstream analysis as shown by Aibar et al.[17]. Inspired by this approach, the regulon activity matrix was used in place of one produced by principal component analysis to perform UMAP projection. Given that regulon activity is measured using an ensemble of genes instead of individual genes, the regulon activity-based clustering is more robust to dropout[17]. Figure 1.5c right panel presents the UMAP projection of heart cells based on the regulon activity matrix. In this plot, it is clear that the regulon activity captured the underlying differences in the cellular population.


Double negative T cell
CD4 Memory
CD8 Naive
CD4 Naive
CD8 effector
pre-B cell
B cell progenitor
NK cell
pDC
CD16 + Monocytes
CD14 + Monocytes

(a)


CD8 effector
CD8 Naive
CD4 Memory
CD4 Naive
NK cell
CD14+ Monocytes
CD16+ Monocytes
Double negative T cell
pDC
pre-B cell
B cell progenitor
Identity
PAX5
IRF7
NR1D1
TCF7L2
CEBPD
RUNX3
LEF1
TCF7
BCL11B
Features
Average Regulon Activity
3
2
1
0
-1

(b)


B cell progenitor
CD4 Memory
CD4 Naive
CD8 Naive
CD8 effector
CD14+ Monocytes
CD16+ Monocytes
Double negative T cell
NK cell
pDC
pre-B cell

(c)

Figure 1.5: Regulon activity inference.

(a) The regulon activity heatmap is plotted using hierarchical clustering. Y-axis represents regulon names, and X-axis represents the PBMC cell types. The regulon activity is translated to color intensity (with blue being the minimum and red being the maximum activities). (b) Navigating the regulon activity of some well-characterized TFs across PBMC cells. Y-axis represents the annotated clusters, and X-axis represents regulon names. The color encodes the average regulon activity across the cluster. (c) (left) UMAP projection of single-cell RNA based on principle component analysis. Each color represents a cell type explained in the legend. (right) UMAP projection of single-cell RNA based on the TF regulon activity. Each color represents a cell type described in the legend

I further analyzed PAX5 target genes detected by the sci-GRN framework. Table 1.2 presents the top 10 targets of the PAX5 gene. Gene Ontology (GO) enrichment analysis was performed on the ranked targets of the PAX5 transcription factor. Table 1.2 presents the top detected biological processes and reveals many B-cell-specific processes. To further test the significance of the ranked targets in detecting relevant biological processes, PAX5 targets were randomly shuffled and used for GO analysis. No biological processes were detected using the randomly ranked targets.

Finally, the top 10 targets of PAX5 detected by sci-GRN were compared to the top 10 targets of PAX5 with no GRN base provided (Table 1.3). For each predicted interaction, PAX5 ChIP-seq from the ENCODE database was used to validate the interaction. ChIP-seq validated all 10 out of 10-top targets of sci-GRN, and 4 out of 10 targets of GENIE3 inference without the GRN base (Table 1.3).

target	importance	ENCODE
CD79A	0.076	yes
TNFRSF13C	0.064	yes
CD22	0.06	yes
SWAP70	0.046	yes
AFF3	0.045	yes
CD72	0.043	yes
CD40	0.039	yes
BLK	0.038	yes
POU2AF1	0.034	yes
CDCA7L	0.033	yes

Table 1.1: Top 10 targets of PAX5 obtained by sci-GRN

The target column represents the PAX5 targets inferred by sci-GRN. The importance column represents GENIE3 feature importance. The Encode column indicates if the target has been identified by the ENCODE consortium.

description	FDR q-value	enrichment
B cell activation	3.52E-02	7.93
lymphocyte activation	1.76E-02	7.93
leukocyte activation	3.18E-02	7.14
cell activation	1.27E-01	5.95
regulation of B cell proliferation	1.21E-01	6.61
regulation of mononuclear cell proliferation	1.01E-01	6.61
regulation of lymphocyte proliferation	8.65E-02	6.61
regulation of B cell activation	7.57E-02	6.61
immune system process	1.44E-01	3.66
adaptive immune response	1.67E-01	39.67
regulation of leukocyte proliferation	2.04E-01	8.50

Table 1.2: GO enrichment analysis using PAX5 ranked targets

GO enrichment using PAX5 ranked targets. The description column contains the name of the identified biological process.

target	importance	ENCODE
IGHM	0.057	no
CD79B	0.055	yes
ABCG2	0.053	no
PLD5	0.052	no
CD79A	0.051	yes
IGHD	0.049	no
MS4A1	0.048	yes
IGKC	0.044	no
BMP3	0.043	no
TNFRSF13C	0.043	yes

Table 1.3: PAX5 target genes obtained by only using GENIE3 GRN inference (without providing a base)

The target column represents the GENIE3's predicted PAX5 targets. The importance column indicated the GENIE3 feature importance. The ENCODE column indicates if the interaction has been recorded in ENCODE consortium.

I next describe an application of applying these techniques to inferring a GRN for the developing human heart using data on individual cells from both RNA-seq expression data and ATAC-seq chromatin data. I first provide a brief background on the major anatomy of the heart that is pertinent to this study

1.5 Results

1.5.1 Single-cell heart development dataset

The human heart is one of the first organs to form during embryogenesis[81]. By week 3, it is a linear heart tube, and between week 3 and 8, it undergoes a complex looping process to form the heart's four chambers, the left atrial (LA), right atrial (RA), left ventricular (LV), and right ventricular (RV), in addition to forming the interventricular septum (IVS). Between week 8 up until birth, the heart undergoes a period of rapid growth and expansion as the various cardiovascular lineages mature to support the developing fetus (Figure 1.6). Defects in this process lead to congenital heart defects, which affect 0.8% to 1.2% of of all live births[82]. Therefore, understanding heart development will give insight into the mechanisms leading to the disease, and principles can further be applied for regenerative medicine purposes.

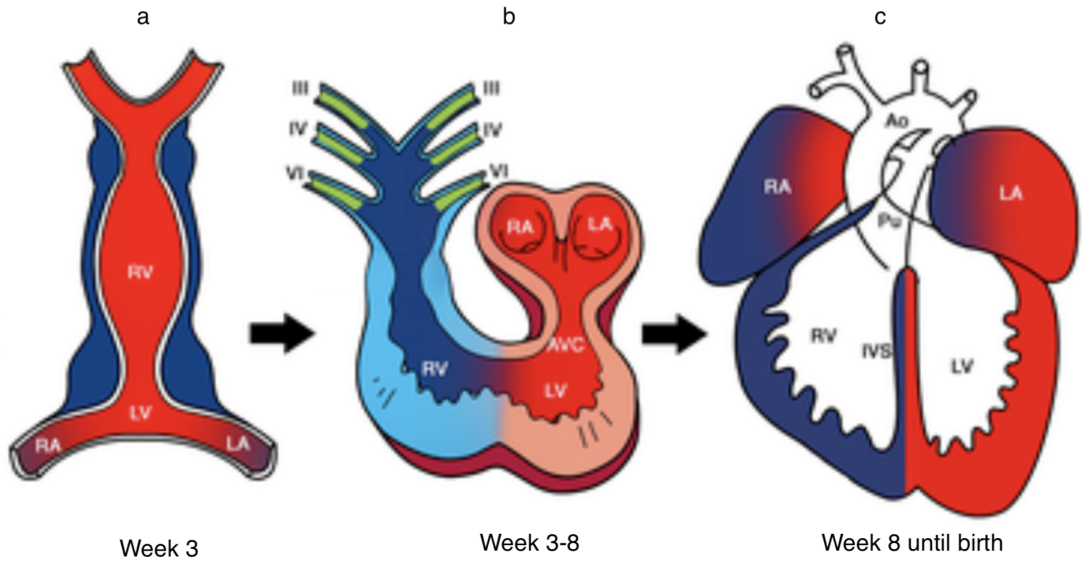


Figure 1.6: Stages of cardiac morphogenesis (image borrowed from Basu et al. 2018[1]

Key stages of human heart embryogenesis. (a) At week 3, human heart is a linear tube. (b) From week 3 to 8, the four heart chambers (LA, LV, RA, RV) and the IVS region are formed. From week 8 up until birth, the heart goes under expansion and enlargement.

For this study, heart single-cells were generated from week 15 of development. Each region of the heart was dissected and divided into two replicates. Two samples from each replicate were sent for scRNA-seq and scATAC-seq (Figure 1.7).

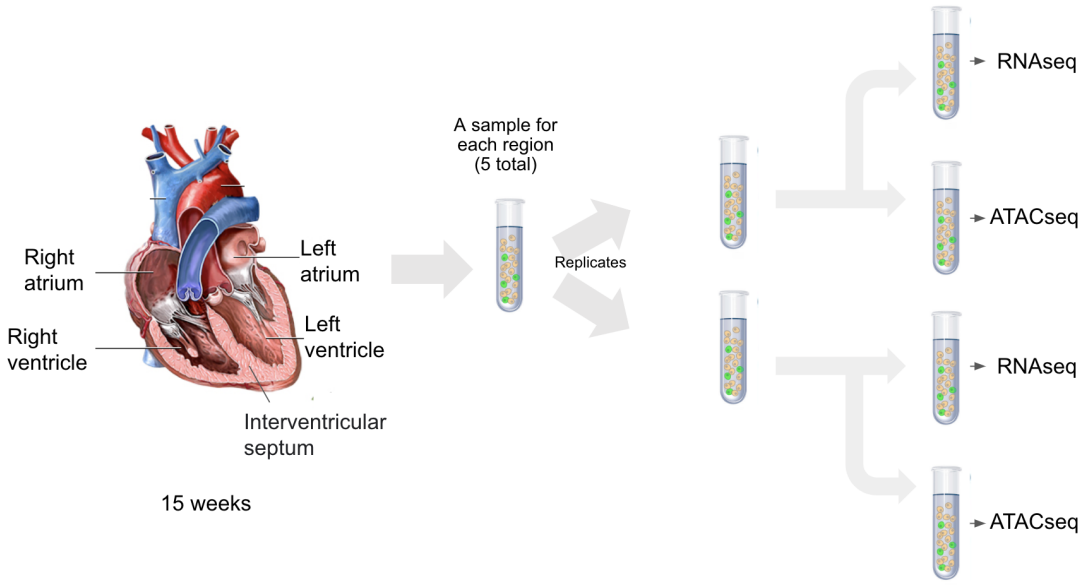


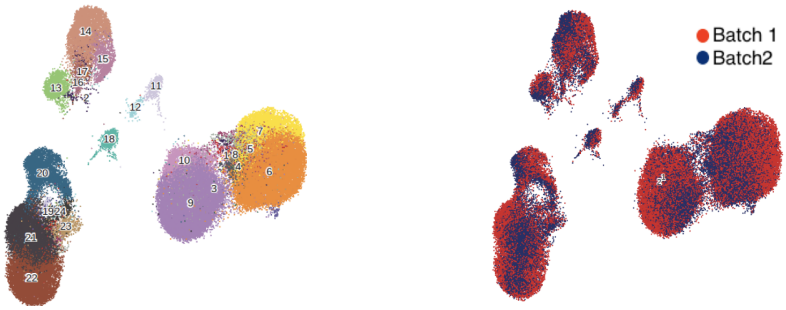
Figure 1.7: Heart sample generation

The heart was dissected into 5 regions (LV, RV, LA, RA, and IVS). Each region's sample was divided into two replicates. Each replicate was split into two; one went for scRNA-seq and the other from scATA-seq.

scRNA sequencing was performed on the collected heart data using 10X Genomics platform (see Methods, Section scRNA-seq processing). For the cardiomyocyte dataset, the top 3000 variable genes were retained and 50 principle components were found to be significant (determined by ElbowPlot).

The heart scATAC sequencing was performed using 10X Genomics chromium platform (see Methods, Section scATAC-seq processing). ArchR[60] was used to preprocess ATAC cells. 81,786 cells passed the quality control step, which included removing cells with a low number of fragments and dead cells[60]. 24 clusters were identified using

ArchR's iterative-LSI method with the default parameters[60] (Figure 1.8a). Batch effects were examined across the two cardiac batches, and the umap projection indicated that iterative-LSI successfully removed a batch effect between the two replicates (Figure 1.8b).



(a) **(b)**

Figure 1.8: scATAC clustering using ArchR's iterative-LSI

(a) Heart scATAC data UMAP projection. 24 clusters were detected in heart scATAC data using ArchR's iterative-LSI method. Each cluster is labeled with a color and number. (b) Batch effects visualization after iterative-LSI clustering. Each batch is color-coded with batch 1 as red and batch 2 as blue.

1.5.2 Single-cell RNA and single-cell ATAC integration

Nine cell types were confidently detected in scATAC-seq data by CCA integration: aCM, vCM, fibroblast, SMC, WBC, endocardial, epicardial, neuronal-like, and VE (Figure 1.9a). RBC, pacemaker-like, and Unknown-CM cell types were not detected in the scATAC-cell population. This could be due to the low number of these cell types in scATAC data, the low number of these cells in scRNA (for example, there are only

31 RBC in the scRNA (Table 1.4)), or the high similarity of these cell types to other cell types (pacemaker cells are similar to atrial cells, and Unknown-CM are similar to cardiomyocyte cells). The annotated LEC cells showed poor performance (evaluated by LEC marker gene activity explained in the next paragraph) and were disregarded from downstream analysis.

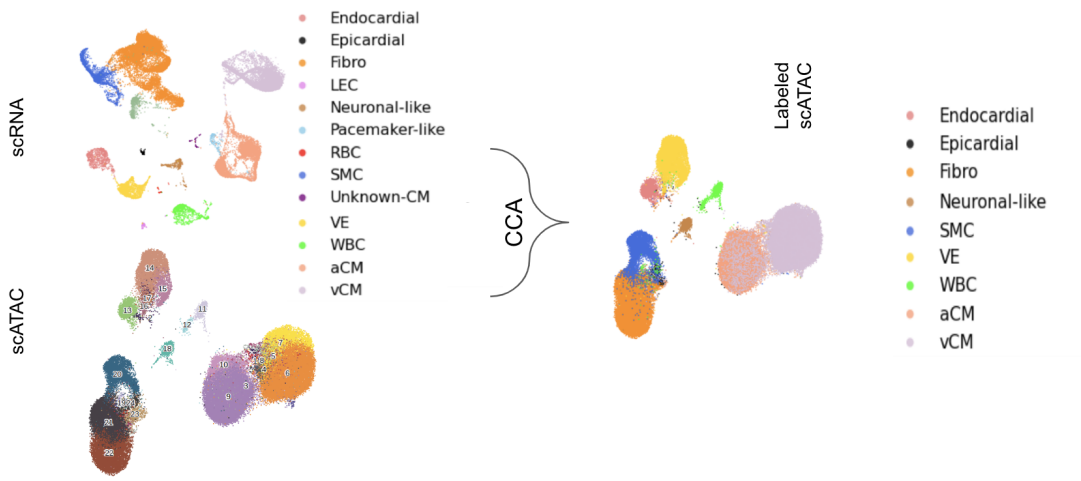
The inferred expression and cell type of each ATAC cell were validated using cell type-specific marker genes (Figure 1.9b). NPPA is a known marker gene for atrial cardiomyocytes and is essential for regulating energy metabolism, vascular remodeling, and cardiac function[83]. Variants in NPPA gene have shown to be associated with cardiovascular diseases[83]. The inferred expression of NPPA (Figure 1.9b) corroborates the scRNA expression of NPPA gene in atrial (Figure 1.9c). MYL2 is a known marker gene of ventricular cardiomyocites[84, 85], and is essential for cardiac muscle contraction[86]. The inferred expression of MYL2 (Figure 1.9b) indicates a higher expression of this gene in ventricular cardiomyocytes, which concurs with the scRNA-seq expression of MYL2 in ventricular cardiomyocytes (Figure 1.9c). TCF21 is a marker gene for fibroblast cells and has been reported to play essential roles in fibroblasts development in the heart[12]. Variants in TCF21 have been reported to be associated with congenital heart diseases[12, 87, 88]. The inferred expression of TCF21 (Figure 1.9b) corroborates the scRNA expression of TCF21 gene in fibroblast (Figure 1.9c). FOXC1 is a marker gene for endocardial cells and is critical for normal heart development[89]. Mutations in FOXC1 locus have been associated with congenital heart defects[90, 89]. Both inferred expression and scRNA demonstrated a higher expression of FOXC1 in en-

docardial cells (Figure 1.9b, Figure 1.9c). WT1 is critical for epicardial cells maturation and shape[91]. WT1 inferred expression and scRNA expression indicate a high expression of WT1 in endocardial cells (Figure 1.9b, Figure 1.9c). MYH11 is a marker gene of smooth muscle cells, [92, 93], and is essential for maintaining vascular wall stability[94]. Figure 1.9b demonstrates a high inferred expression of MYH11 in smooth muscle cells. CD14 is a marker gene for immune cells[95], and the inferred expression demonstrates a high CD14 expression in white blood cells (Figure 1.9b). CLDN5 regulates blood vessels permeability[96], and is highly expressed in vascular endothelial (Figure 1.9c). The inferred expression of CLDN5 corroborates the scRNA expression (Figure 1.9b). Litviňuková et al. reported MPZ as a marker gene of neuronal-like cells in the heart[97]. The inferred expression of MPZ in scATAC cells and the scRNA data corroborate their finding (Figure 1.9b).

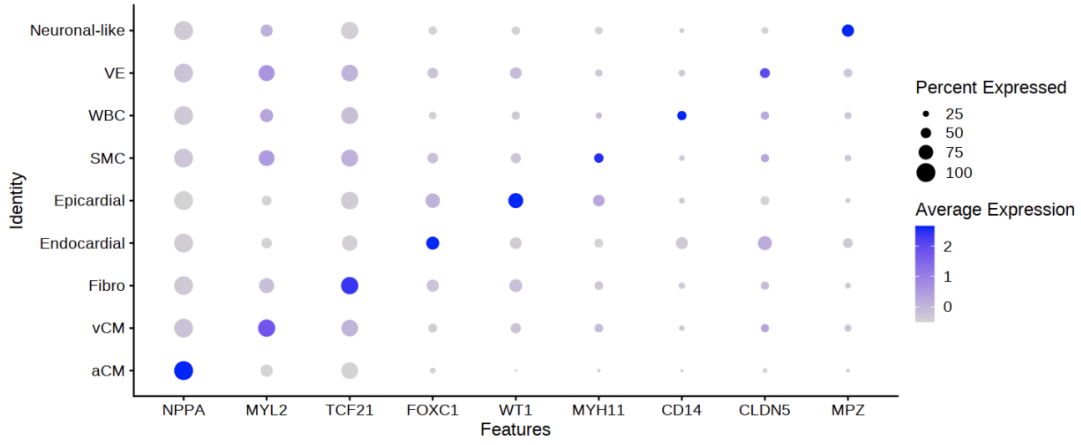
count	cell type
31	RBC
1,090	WBC
233	LEC
1,826	VE
863	Endocardial
598	Neuronal-like
129	Epicardial
12,099	Fibro
2,064	SMC
62	Unknown-CM
223	Pacemaker-like
6,132	aCM
6,369	vCM

Table 1.4: Annotated scRNA-seq cell types

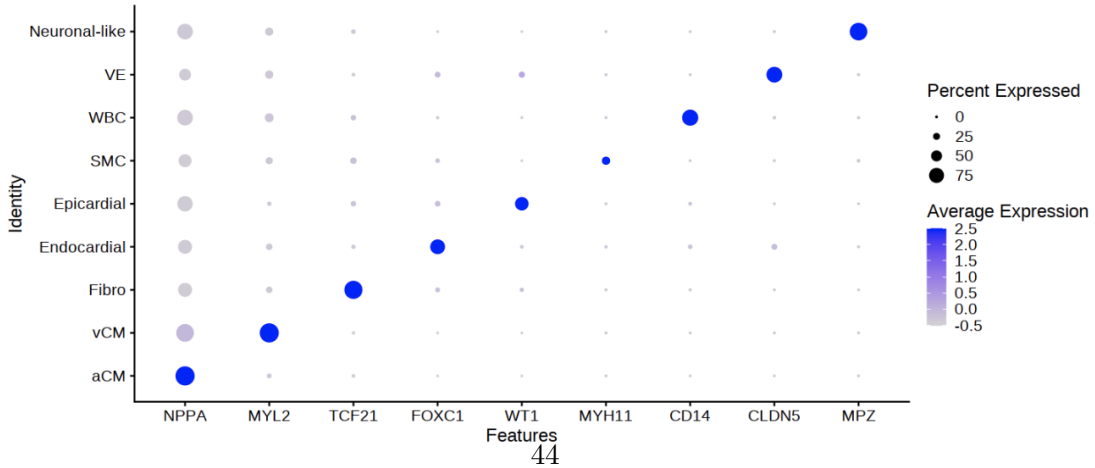
The number of cells in each annotated scRNA cluster


scRNA
scATAC
Endocardial
Epicardial
Fibro
LEC
Neuronal-like
Pacemaker-like
RBC
SMC
Unknown-CM
VE
WBC
aCM
vCM
CCA
Labeled scATAC
Endocardial
Epicardial
Fibro
Neuronal-like
SMC
VE
WBC
aCM
vCM

(a)


Neuronal-like
VE
WBC
SMC
Epicardial
Endocardial
Fibro
vCM
aCM
Identity
NPPA
MYL2
TCF21
FOXC1
WT1
MYH11
CD14
CLDN5
MPZ
Features
Percent Expressed
25
50
75
100
Average Expression
2
1
0

(b)


Neuronal-like
VE
WBC
SMC
Epicardial
Endocardial
Fibro
vCM
aCM
Identity
NPPA
MYL2
TCF21
FOXC1
WT1
MYH11
CD14
CLDN5
MPZ
Features
Percent Expressed
0
25
50
75
Average Expression
2.5
2.0
1.5
1.0
0.5
0.0
-0.5

(c)

Figure 1.9: Human heart scRNA and scATAC integration.

(a) Inferring the cell type and expression of each ATAC heart cell through scRNA and scATAC integration. The left-top UMAP projection represents annotated heart scRNA-seq data. The left bottom UMAP projection represents the heart scATAC-seq data clustering obtained by the iterative-LSI. The right UMAP represents the annotated scATAC-seq data. (c) Navigating the expression of marker genes across scRNA cell types using Dot plot. Y-axis represents the annotated clusters, and X-axis represents the marker genes. The size of each dot represents the percentage of cells within the cluster expressing the marker gene. The color encodes the average expression of each marker gene across the cluster. (b) Navigating the inferred expression of marker genes across scATAC cell types using Dot plot. Y-axis represents the annotated clusters, and X-axis represents the marker genes. The size of each dot represents the percentage of cells within the cluster expressing the marker gene. The color encodes the average expression of each marker gene across the cluster.

1.5.3 Detecting enhancer dependent regulations

Table 1.5 provides a summary of the number of interactions/edges that GENIE3 needs to predict without providing a GRN base vs. when a GRN base is provided. As shown in Table 1.5, the number of interactions are narrowed down from 20,459,790 to 3,816,362 after providing a GRN base. sci-GRN implicates specific binding sites as actively bound by a TF to regulate a target gene's activity. Therefore, the sci-GRN framework allows finding the TF-bound regions important for regulating a gene's expression. I tested the capability of the sci-GRN framework in detecting a recently discovered enhancer in the fetal heart. Man et al. discovered a novel enhancer

region around 40kb upstream of two atrial-specific genes, NPPA and NPPB, named RE1[98]. They demonstrated that RE1 deficient mice had enlarged hearts before and after birth[98]. Furthermore, they showed that RE1 enhancer is occupied by TBX5 transcription factor, an atrial-specific transcription factor. Figure 1.10 illustrates the two TBX5 binding site peaks detected by cis-correlation and trans-correlation analysis that overlap the RE1 enhancer. As part of sci-GRN workflow, cis-correlation is used to build the GRN base, which means that one can always retrieve TF-bound genomic regions that regulate target gene's expression. Thus, sci-GRN provides a framework for discovering novel cell-type-dependent regulatory connections and forming hypotheses for wet-lab experiments.

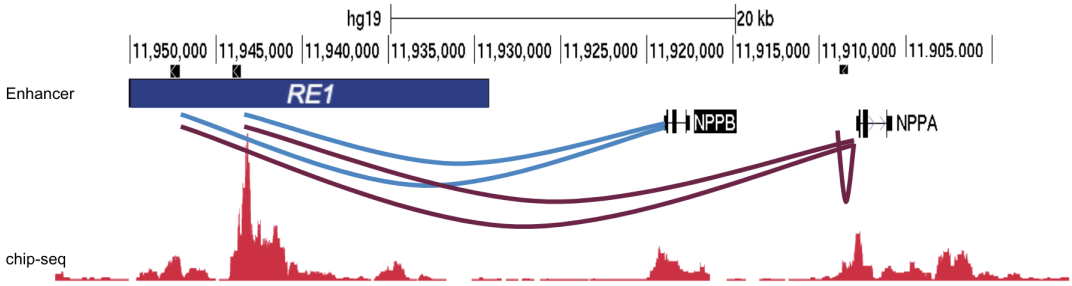


Figure 1.10: RE1 enhancer in proximity of atrial specific genes, NPPA and NPPB
Three peaks with TBX5 binding site in proximity of NPPA-NPPB genes. The two left peaks in the RE1 region are found to be correlated to the inferred expression of NPPA and NPPB by the GRN base reconstruction. The right peak resides in NPPA promoter and is correlated to the NPPA expression based on cis-correlation analysis.

	Without GRN base	With GRN base
Number of genes	23,517	8,462
Number of TFs	870	451
Num. edges	20,459,790	3,816,362

Table 1.5: Heart GRN base summary

Comparing the number of genes, TFs, and edges in sci-GRN (with GRN base) vs. when no base is provided for GENIE3.

1.5.4 TBX5 target genes

TBX5 is an essential regulator of heart development[99] and plays a critical part in regulating atrial rhythm[99, 100]. A genome-wide association study (GWAS) by Holm et al. revealed a strong link between TBX5 locus alteration and atrial fibrillation[101]. Furthermore, studies have shown that mutations in TBX5 are associated with Holt–Oram syndrome, which causes atria conduction dysregulation and upper limb malformations[102, 100, 99].

Table 1.6 represents the top targets of TBX5, inferred by sci-GRN. sci-GRN was able to retrieve well-studied TBX5 target genes, such as NPPA[98, 103] and MYH6[104, 102] (Table 1.6). The NPPA gene encodes for atrial natriuretic factor (ANF), which is highly expressed in the atria and ventricle during embryonic and fetal stages[105]. After birth, NPPA expression is strongly downregulated in the ventricle. Upon stress induction in cardiac cells, the NPPA transcription is drastically increased in both atria

and cardiomyocytes[105]. Because of this feature, the ANF abundance in cardiomyocytes is used to assess heart failure and cardiovascular diseases[105]. MYH6 encodes for myosin heavy chain 6, which plays an important role in myofibril assembly and heart development[106]. Mutations in MYH6 locus have shown to be associated with congenital heart disease, atrial septal defect[107], hypertrophic cardiomyopathy (HCM)[108], and systolic dysfunction[108].

Gene Ontology (GO) enrichment analysis was performed on the ranked targets of TBX5 (371 total) using GOrilla[109, 110]. The GO analysis detected atrial processes, such as heart contraction, blood circulation, muscle contraction, striated muscle contraction, and heart rate regulation (Table 1.7). The significance of the ranked targets on the GO enrichment outcome was assessed by randomly shuffling the TBX5 targets and feeding that to the GO analysis (371 randomly ranked TBX5 targets). Table 1.8 demonstrates that random ranking drastically decreases the significance of detected biological processes, with none being significant or making biological sense. I further tried to assess TBX5 targets using SCENIC (default parameters) but the TBX5 regulon was not detected by the SCNIEC protocol. In addition to TBX5, the targets of a non-cardiac TF, POU5F1, were used for GO enrichment analysis. GO enrichment performed on POU5F1 targets determined no significant biological processes.

target	importance	evidence
MYH6	0.071	Jiang et al., 2020
NPPA	0.062	Mori et al., 2006
EPHA4	0.052	-
NR2F1	0.051	-
PRKAG2	0.046	Kathiriya et al., 2021, suggests that PRKAG2 is a TBX5 target
KCNA5	0.04	Mori et al., 2006, suggests that KCNA5 is a TBX5 target
VSTM2L	0.039	-
CPNE5	0.037	Mori et al., 2006, suggests that CPNE5 is a TBX5 target
ADCY5	0.032	-
CORIN	0.032	Pan et al., 2002

Table 1.6: Top 10 targets of TBX5 predicted by sci-GRN

The target column represents the target gene name. The importance column represents the random forest importance value. The evidence column represents the studies that support the TF-target interaction.

Biological Processes	FDR q-value	Enrichment
regulation of system process	1.89E-06	2.69
regulation of heart contraction	1.18E-05	2.94
regulation of blood circulation	4.1E-05	2.83
muscle contraction	1.64E-04	3.14
tissue morphogenesis	1E-03	2.82
regulation of membrane potential	1.24E-03	3.51
striated muscle contraction	1.15E-03	3.34
regulation of transport	5.17E-03	2.47
regulation of heart rate	8.97E-03	3.08
muscle system process	9.93E-03	2.58

Table 1.7: Ranked-base GO enrichment analysis using TBX5 target genes

Biological Processes	FDR q-value	Enrichment
regulation of telomere maintenance	1E+00	47.20
regulation of telomere maintenance via telomerase	1E+00	47.20
regulation of telomere maintenance via telomere lengthening	1E+00	47.20
regulation of phenotypic switching	9.36E-01	32.18
regulation of smooth muscle cell differentiation	7.49E-01	32.18
regulation of vascular smooth muscle cell differentiation	6.24E-01	32.18

Table 1.8: GO enrichment analysis using a random ranking of TBX5 targets

1.5.5 Regulon activity inference

The univariate linear model (ULM)[111] implemented in decoupleR package[25] was used to infer regulon activity for 427 transcription factors. ULM uses the predicted GRN from sci-GRN as the input to a univariate linear model to model scRNA count matrix [111]. The t-values from the fitted model represent the regulon activity[25]. To facilitate further statistical analysis, a statistical threshold could be derived from a regulon activity to group cells into "on" and "off" states. The regulon activity state in a cell could be determined based on bimodal regulon activity distribution[17]. As described in the SCENIC paper[17], a two-component Gaussian mixture model was fitted to each inferred regulon activity to obtain a threshold[17]. Figure 1.11 represents the thresholds for the on/"off" states between atrial and ventricular cardiomyocytes cells, using three well-characterized cardiac atrial cardiomyocytes, TBX5, NR2F1, and NR2F2. For each of these transcription factors, the inferred regulon activity threshold separated the two cell types. In theory, cells to the left of the dotted line are in an "off" state, and cells to its right are in an "on" state, corroborating with cardiac biology.

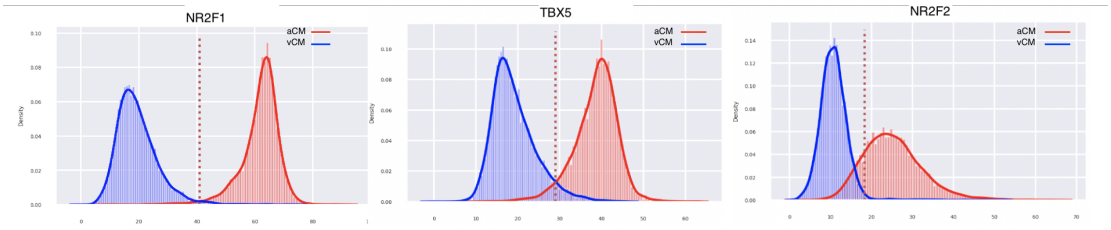


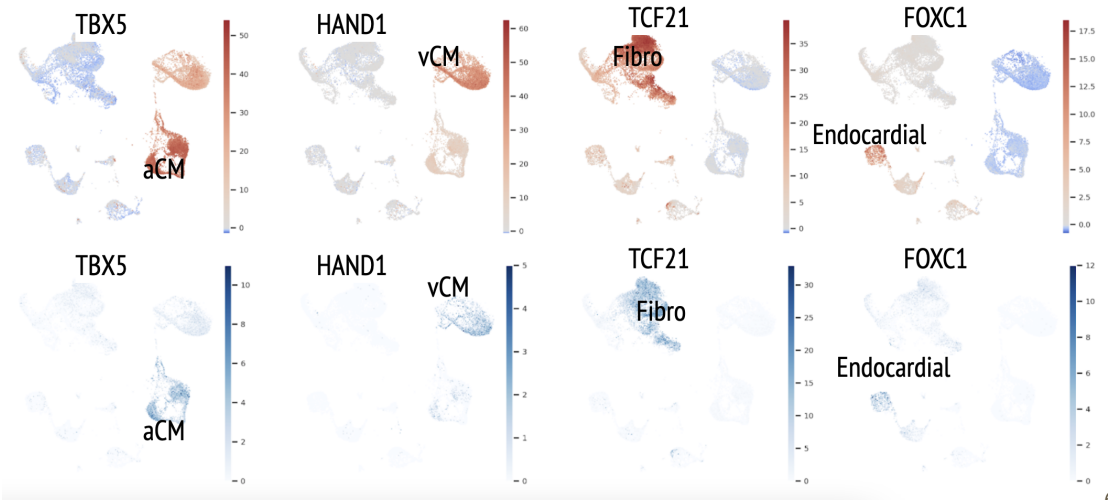
Figure 1.11: Regulon activity binarizartion

The regulon activity distribution of NR2F1, TBX5, and NR2F2, across atrial (red) and ventricular (blue) cells. The dotted line represents the inferred threshold based on the two-component Gaussian mixture model. Cells to the right of the dotted line are in an "on" state, and cell to the left of the dotted line are in an "off" state.

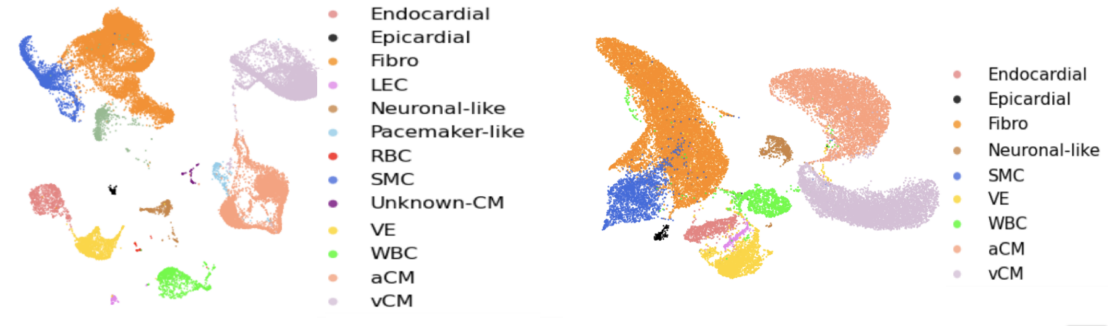
The activity of some well-known transcription factors (regulon activity) was visualized across the heart cells. Figure 1.12a shows the inferred activity of four of these regulons TBX5, HAND1, TCF21, and FOXC1. TBX5 is a well known atrial regulon. During early fetal heart development, TBX5 mainly acts as a transcription factor regulator of cardiomyocyte maturation genes[99]. In the later stages of cardiac development, TBX5 is involved in regulating and maintaining the cardiac conduction system[99, 112, 113]. Figure 1.12a demonstrates high TBX5 activity in the atrial population. Next, I looked at HAND1 regulon activity, which is a ventricular-specific transcription factor[114]. Mutation in HAND1 has been shown to contribute to diastolic heart failure[114]. Figure 1.12a shows a high activity of HAND1 regulon in ventricular cardiomyocytes. Furthermore, I assessed TCF21 regulon activity, which is essential for cardiac fibroblast formation[115], and observed high activity in the fibroblast cluster.

Finally, I tested the activity of FOXC1, an endocardial-specific transcription factor that plays an essential role in cardiac outflow tract morphogenesis[90, 116], and observed high activity in the endocardial cluster.

The regulon activity matrix was used for downstream analysis as was shown by the work of Aibar et al.[17]. For example, one can use this lower dimensionality matrix instead of principal component analysis to perform UMAP projection. Given that regulon activity is measured as a whole instead of individual genes, the regulon activity-based clustering is more robust to dropout[17]. Figure 1.12b right panel presents the UMAP projection of heart cells based on the regulon activity matrix. The regulon activity-based clustering presented in Figure 1.12b demonstrates that the heart regulon activity matrix captures the underlying differences in the cellular population as it successfully separates cell types.



(a)



(b)

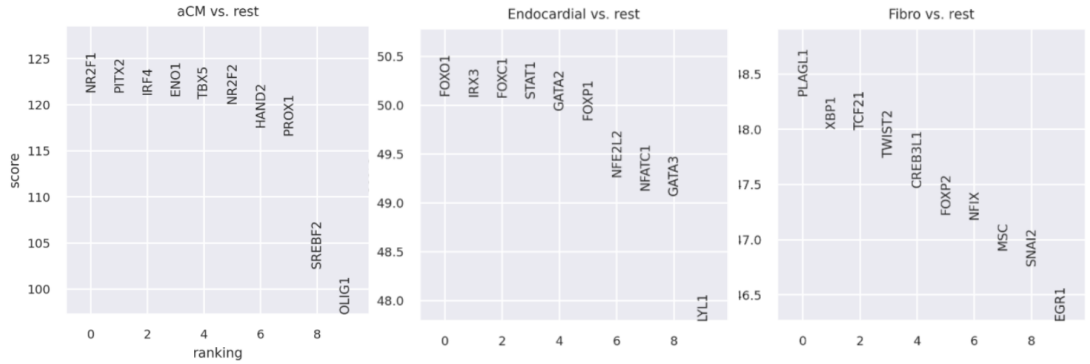
Figure 1.12: Regulon activity inference.

(a) (top) Inferred regulon activity of 4 cardiac transcription factors, TBX5, HAND1, TCF21, and FOXC1 across heart cells. Color intensity from blue-red is used to present regulon activity. (bottom) The RNA expression of the corresponding transcription factors. Blue intensity color is used to present regulon activity. (b) (left) UMAP projection of scRNA data based on principal component analysis. (right) UMAP projection of single-cell RNA based on the TF-activity. The cell type annotation is provided in the legends.

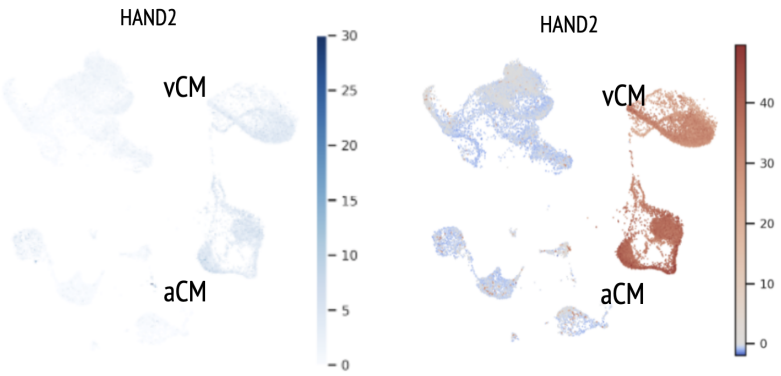
To explore regulon activity specificity across heart cell types, Wilcoxon rank-

sum statistical test[117] was used. Figure 1.13a presents the most substantial regulons in atrial, endocardial, and fibroblast cells. As expected, FOXC1 demonstrated high specificity for endocardial, and TCF21 demonstrated high specificity for fibroblast cells. I further looked at regulons with the highest activity in atrial cardiomyocytes and validated the high activity of the following regulons through literature review: (NR2F1[118], PITX2[119], TBX5[103], NR2F2[120], PROX1[121]). Furthermore, HAND2 was found to have high specificity for atrial cardiomyocytes (Figure 1.13a). Figure 1.13b presents high HAND2 expression, and Figure 1.13b shows high HAND2 regulon activity in atrial cardiomyocytes.

Little is known about the vital role of HAND2 in atrial cardiomyocytes. Our integrative analysis suggests that HAND2 could potentially be consequential in atrial cardiomyocytes. Our finding corroborates the TBX5 haploinsufficiency study by Kathiriya et al. Kathiriya et al. conducted TBX5 dosage-dependent experiments across the human iPSC-derived cardiomyocyte developmental stages and investigated the effects of TBX5 dosage changes in GRNs inference. Their results indicated a decrease in HAND2 pagerank centrality, suggesting that it might be a TBX5 target. Further wet-lab experimental validations are needed to corroborate the specificity of HAND2 regulon in the atria.



(a)



(b) **(c)**

Figure 1.13: The regulon activity specificity

(a) Ranking of the differentially active regulons across heart aCM, endocardial, and fibroblast clusters using Wilcoxon rank-sum. The Y-axis represents the Wilcoxon rank-sum statistical test score, and the X-axis represents the regulon ranking. (b) HAND2 expression is translated into a range of blue colors based on the amount expressed in each cell. (c) HAND2 regulon activity is translated into a range of blue-red colors based on the HAND2 regulon activity in each cell.

1.6 Conclusion

Building a valid Gene Regulatory Network (GRN) could help scientists generate novel biological hypotheses about molecular interactions, design perturbation experiments, and use networks as biomarkers[52], as they function in organogenesis, tissue maintenance, and disease.

Most single-cell GRN inference methods build the GRN using only the single-cell RNA data[11, 12]. Popular methods such as GENIE3[15] rely on single-cell RNA to infer GRNs. The challenge such methods face is that correlation in expression does not necessarily mean that a gene is a direct target of a transcription factor since many genes can have correlated expression through indirect causes. Methods such as SCENIC[17] tackle this problem by pruning indirect targets (edges) by looking at the transcription factor (TF) enrichment around a transcription start site. Such an assumption reduces the false discovery rate, but at the same time, disregards the existence of any distal-dependent enhancers.

Here, I tackle this issue by building GRNs through scRNA and scATAC integration. The sci-GRN workflow I developed consists of a) Inferring the cell type and expression of each ATAC cell, b) creating initial GRN (the GRN base) through cis-interaction and trans-interaction correlation analysis, and c) Ranking TF-target interactions through random forest regression (GENIE3). I benchmarked sci-GRN workflow on PBMC dataset and demonstrated its application to a novel human heart dataset.

Investigation of positive control TFs in the PBMC dataset (obtained from 10X

genomics site[71]) revealed that the inferred regulon activities corresponded well with the published literature. Furthermore, The top 10 targets of PAX5 TF (a well-known B-cell marker) identified by sci-GRN were validated using ENCODE database. Comparing the PAX5 top 10 targets identified by sci-GRN vs. the GENIE3 identified targets (without a GRN base provided) indicated more evidence of PAX5-target interaction in sci-GRN (using ENCODE ChIP-seq database). Further GO enrichment analysis using PAX5 targets identified by sci-GRN determined B-cell-specific biological processes.

The sci-GRN predicted GRN base defines specific binding sites as actively bound by a TF to regulate a target gene's activity. Therefore, sci-GRN framework allows uncovering the TF-bound regions important for regulating a gene's expression. Investigation of NPPA regulatory regions bound by TBX5 TF using the sci-GRN predicted GRN base identified a recently discovered cardiac enhancer, RE1. Thus, sci-GRN provides a framework to identify regulatory regions important for gene regulations. The regulon activity analysis of several well-characterized cardiac TFs revealed that the inferred regulon activities corresponded well with the published literature. GO enrichment analysis on the ranked targets of TBX5 identified atria-specific processes, corroborating the heart biology. Compared to sci-GRN, SCENIC did not detect the TBX5 regulon (default parameters were used). This could be due to the conservative nature of SCENIC workflow and its limited searching space (+/-10kbp of TSS).

Finally, the sci-GRN framework predicted a high HAND2 regulon activity in atrial cardiomyocytes. The predicted HAND2 regulon requires further wet-lab experiment validation. For example, the HAND2 ChIP-seq assay could be conducted in

atrial cardiomyocytes to show that it binds to enhancers near (or regulating) atrial-related genes. Another experimental validation would be looking at atrial cells (mainly cardiomyocytes) in HAND2 knockout vs. wild-type to see if there is a difference in specification or maturation of the atrial cardiomyocytes (developmental defects).

Chapter 2

Impact of anti-obesity medication initiation and duration on weight loss in a comprehensive weight loss program

The text of this chapter includes reprints of the previously published material. Safavi R, et al. (2019). My contribution to this study involves statistical evaluation of anti-obesity medication initiation, usage, and duration on weight loss of patients enrolled in Enara Health's weight loss program.

Obesity Science & Practice doi: 10.1002/osp4.361

ORIGINAL ARTICLE

Impact of anti-obesity medication initiation and duration on weight loss in a comprehensive weight loss programme

R. Safavi[1], A. Lih[2], S. Kirkpatrick[2], S. Haller[2] and M. R. Bailony[2]

[1]University of California, Santa Cruz, California, USA; [2]Research Department, Enara Health Inc., San Mateo, California, USA;

Received 19 April 2019; revised 1 July 2019; accepted 3 July 2019

Correspondence to: MR Bailony, Research Department, Enara Health Inc., 1650 S Amphlett Blvd, Ste 115, San Mateo, CA 94402, USA.
E-mail: rami@enarahealth.com

Summary

Objective

This retrospective study aimed to evaluate the impact of anti-obesity medication (AOM) initiation, usage and duration on weight loss in a 72-week precision obesity programme. The type of AOM, diet and exercise plan was chosen based upon an individual's biological and psychosocial needs. The 72-week study duration allowed for a fair investigation of the downstream impact of delayed versus early AOM initiation.

Methods

Participants, aged ≥18 years with body mass index ≥30 kg m^{-2}, enrolled from 1 March 2015 to 1 April 2017, were included. Subgroups were assigned by AOM usage (users versus non-users, early [before 8 weeks] versus delayed [after 8 weeks] AOM initiation and short [<6 months] versus long [≥6 months] AOM duration). Primary endpoints included change in baseline weight at 72 weeks and proportions achieving ≥5%, ≥10% and ≥15% weight loss. Outcomes were compared between subgroups.

Results

Mean age and body mass index (N = 129) were 45.0 ± 14.0 years and 37.0 ± 6.0 kg m^{-2}, respectively; 67% were female. At week 72, AOM users (N = 71) achieved significantly greater mean percentage reduction in baseline weight than non-users (N = 58). On average, baseline weight decreased by 14.04 ± 6.2% in users versus 10.9 ± 6.8% in non-users (P = 0.008); 84% and 94% of non-user and AOM users lost >5% weight loss (P = 0.006). A higher proportion of users lost ≥15% of weight (45.1% vs. 19.0%; P < 0.001). Mean percentage reduction in weight was greater for early versus delayed starters (−17.60 ± 5.3% vs. −13.95 ± 5.5%; P = 0.024), and longer AOM usage trended towards increased weight loss.

Conclusion

Early initiation of AOM may enhance weight loss.

Keywords: obesity treatment, weight loss, weight management programme, weight-reducing drugs.

Introduction

Recent guidelines generally recommend anti-obesity medications (AOMs) only for patients who have failed to achieve 5% weight loss through dietary, exercise and/or lifestyle modification interventions alone (1–5). However, the optimal time to introduce an AOM in the setting of a comprehensive weight loss programme has not been well studied (6). Should physicians always start with diet and exercise before adding a medication or should they initiate both together? Does the intensity of diet impact the initiation and efficacy of AOM? How important is medication duration and compliance? As the obesity epidemic continues to grow, the different combinations and order in which therapies are delivered deserve more attention and study (7).

Recently, a study of 150 adults showed that adding liraglutide of 3.0 mg to intensive behavioural therapy

© 2019 The Authors

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

(IBT) for obesity at programme onset showed larger weight loss in the IBT-liraglutide group (11.5 ± 1.3%) than the IBT-alone group (6.1 ± 1.3%) but that meal replacement did not increase weight loss when added to IBT and liraglutide (11.8 ± 1.3%) (8). This differs from a study that showed that of sibutramine, lifestyle modification and a portion-controlled diet produced significantly greater weight loss (16.5 ± 8.0%) than both the sibutramine alone (4.1 ± 6.3%) and the sibutramine plus lifestyle therapy (10.8 ± 10.3%) cohorts (9). These studies suggest that early initiation of medication improves results of lifestyle therapy while dietary intensity may enhance the efficacy of some obesity medications but not others. Restricting the use of AOMs to patients who demonstrate lifestyle failure may not be in best interests of patients when the sum may be greater than the parts.

On the contrary, there is also evidence suggesting that adding AOM only after patients achieve 5% or greater weight loss may produce better long-term weight loss and maintenance. In one study, 561 patients (mean body mass index [BMI], 34.8 kg m^{-2}) were randomized to either of two topiramate (TPM) regimens or placebo after losing ≥8% of their initial body weight following an 8-week low-calorie diet (LCD) (10). At week 44 (including the 8-week LCD run-in), a modified intent-to-treat analysis showed that TPM significantly reduced weight by 15.4% and 16.5% in the 96 and 192 mg d^{-1} arms, respectively, compared with 8.9% in the placebo arm (10). In contrast, TPM randomization from the onset of a weight loss programme has been associated with a 5–6.5% reduction in baseline weight at 52 weeks (11). Similarly, in the SCALE Maintenance trial, patients (N = 212) who initiated liraglutide of 3.0 mg d^{-1} after losing 6% of body weight during a 4- to 12-week LCD run-in achieved 12.2% weight loss (6.2% from randomization) at 56 weeks compared with 6% weight loss in the placebo arm (N = 210) (12). On the other hand, patients randomized to liraglutide at the programme onset achieved only 8% weight loss at 56 weeks (13). While not comparative, these studies collectively suggest that delaying the start of AOM in the context of a comprehensive weight loss programme is not only beneficial for maintaining weight loss but may also increase the total weight loss achieved. The potential for additive weight loss with more precise timing of pharmacotherapy, therefore, warrants further investigation of the optimal time for initiating AOM.

The impact of AOM compliance and duration on weight loss is also unclear. In the real world, many patients discontinue AOM after reaching their desired weight loss goal. In the clinical trials for naltrexone–bupropion (Contrave®), 42% to 49% of patients dropped out due to medication nonadherence or other reasons related to weight gain (14). In the 2-year SEQUEL extension study for controlled-release phentermine/TPM, 84% (568/676) of patients completed the study, while 16% of patients discontinued due to either nonadherence or unrelated reasons (15). In these studies, dropouts were attributed not only to medication nonadherence but also to nonadherence to lifestyle and behavioural changes. There is only one study to our knowledge comparing shorter versus longer duration AOM use. In that study, longer duration phentermine was associated with greater weight loss up to 2 years after initiation (16).

This study evaluated the impact of medication initiation, usage and duration on weight loss in outcomes in a commercial precision medicine weight loss programme that incorporates clinical visits, a mobile application (mHealth) and tailored nutritional programming.

Methods

Study design

This is a retrospective cohort study analysing the impact of AOM on weight reduction in adults participating in a comprehensive precision weight loss programme (Enara Health). Data were collected from the electronic health records of patients who enrolled and started a weight loss journey during the period from 1 March 2015 to 1 April 2017 (with follow-up through 1 October 2018). Enara Health provided the study investigators with all relevant data following de-identification of the entire dataset. The Hummingbird Institutional Review Board (Needham, MA, USA) approved the study protocol and deemed that the study met the requirements for exemption from the US Department of Health and Human Services regulations for the protection of human subjects (45 CFR 46.101[b][4]).

Weight loss programme interventions

Enara Health is a medical company based in San Mateo, California, that offers a unique hybrid digital and in-person weight loss programme. The Enara Health weight loss programme has been operating since March 2015. By combining mobile technology with healthcare visits, the programme provides patients with personalized physician-driven medical weight loss programmes. Upon enrolment, patients are initially evaluated by a physician or physician's assistant who determines the severity of obesity and the patient's readiness to change and performs a full metabolic workup and examination. Patients engage in one-on-one meetings with a registered dietician on a weekly basis during the first 3 months, once every other week during the next 3 months and then once monthly for 6 months.

© 2019 The Authors
Obesity Science & Practice published by World Obesity and The Obesity Society and John Wiley & Sons Ltd Obesity Science & Practice

Patients also see their physician or physician assistant on average once a month or the first 3 months, and then every 3 months thereafter. Patients may conduct their visits either in-person or via face-to-face video communication. There is a monthly fee of $50 for patients with insurance (most insurances accepted) or $300 without insurance.

Between visits, patients utilize an mHealth application, which delivers tailored educational content, messages of encouragement and support, nutritional and behavioural feedback, a review of meals and feedback by a registered dietician, as well as exercise suggestions, planning and encouragement by a fitness trainer. Patients may choose between an intense, rapid weight loss programme (daily caloric intake: 800–1,000 kcal) or a non-intensive weight loss programme that emphasizes non-processed foods but does not encourage calorie counting. Both options encourage whole, unprocessed, low glycaemic foods. Although all participants are assigned a target exercise goal of 150 min week^{-1}, a personal exercise coach determines the duration of exercise based on the patient's physical ability. Many patients utilize activity trackers such as Fitbit® or Apple Watch® to monitor weekly exercise.

Where appropriate, patients are prescribed medications to assist with weight loss. Prescriptions are driven by the patient's request after attending an educational visit on weight loss medications. AOMs are prescribed in accordance with the US Food and Drug Administration-labelled recommendations; however, in the event of non-coverage, generic equivalents are prescribed, and any off-label usage is discussed with the patient. The choice of medication is ultimately made by the patients and their physicians.

Prescribed weight loss medications include lorcaserin, phendimetrazine, diethylpropion, phentermine/TPM, liraglutide and bupropion/naltrexone. Some patients may be prescribed more than one AOM. Routine follow-up visits with the provider are conducted to ensure that the patient is responding positively to the medication and is not experiencing adverse effects. All patients are started on the lowest dosage of medications, while dose adjustments are made on an individual basis as needed to achieve the desired outcomes.

Study cohort selection

Patients enrolled in the Enara Health weight loss programme during the study period, who were aged ≥18 years with a BMI ≥30 kg m^{-2} at enrolment, were eligible for inclusion. Patients who became pregnant during the study period were excluded from the analyses.

Assignment of anti-obesity medication subgroups

Overall, participants were matched into a cohort so long as their start date indicated that they could have participated for the study duration regardless of their engagement status at that time point. Subgroups were assigned by AOM usage (users versus non-users), timing of AOM initiation (early versus delayed) and treatment duration (short term versus long term) and were defined as follows.

Anti-obesity medication users versus anti-obesity medication non-users

A patient was defined as an AOM user so long as the electronic health records documented medication use for a period exceeding 1 month. All other patients were considered to be non-users. A 1-month cut-off was deemed necessary because many patients who had been prescribed AOM had not initiated treatment due to coverage-related or cost-related reasons, and compliance was only recorded on a monthly basis in the medical record.

Early starters versus delayed starters

For patients who used more than one AOM, the status of initiating an AOM was assigned based on the date of the first AOM prescribed. Early starters were defined as patients who initiated an AOM within 2 months of the date of programme enrolment, whereas delayed starters initiated an AOM >2 months after the date of programme enrolment.

Short duration versus long duration anti-obesity medication

Anti-obesity medication treatment duration represented the period commencing on the date of the first recorded AOM prescription and ending on the date of cessation of all AOMs prescribed. Patients who received ≥1 AOM for a total duration of 1–6 months were considered to have short duration AOM treatment, whereas patients who received ≥1 AOM for >6 months had long duration.

Study data collection

Enara Health collected demographic and clinical data for all study patients at the initial clinic visit. Height, body weight, blood pressure, pulse and body composition were measured at the first appointment to establish baseline values and then rechecked in the clinic at every in-person appointment thereafter. Body weight was

© 2019 The Authors
Obesity Science & Practice published by World Obesity and The Obesity Society and John Wiley & Sons Ltd Obesity Science & Practice

collected daily via the patient's home scale, which connected directly to Enara Health's provider dashboard. Medical providers documented AOM compliance on a monthly basis.

Study endpoints

The intention-to-treat (ITT) duration and time endpoint for the primary analysis were 72 weeks. Primary endpoints included weight change at 72 weeks from baseline and the proportion of patients who lost ≥5%, ≥10% and ≥15% of their baseline body weight. Secondary outcomes included weight change from baseline and categorical weight loss at 24 and 48 weeks.

Statistical analysis

In the primary analysis of weight loss in AOM users and non-users, an ITT analysis was performed. All individuals who started and enrolled in a weight loss programme were included in this analysis. If missing, imputed data were carried through from the last measured observation to the study endpoint. A 'completers' analysis was conducted to evaluate the subgroups characterized by AOM initiation and duration due to the potential bias associated with missing data carried forward and given that AOM status imputations with statistical modelling were underpowered. The completers analysis included all patients who had data on weight loss and AOM status at 72 weeks.

At each time endpoint of interest, change in baseline body weight was expressed as the mean percentage change and the mean (± standard deviation) absolute difference in kilograms of body weight. The proportions of patients achieving ≥5%, ≥10% and ≥15% weight loss in each subgroup were also calculated. Differences in weight change and in the proportions attaining each percentage category of weight loss were compared between the subgroups using a *t*-test for continuous data and a Chi-square test for categorical data.

All statistical analyses were performed using Python 3.6. Statistical significance was defined as $P < 0.05$.

Results

Selection of study cohort

Figure 1 shows the disposition and selection of the final study cohort and AOM subgroups. Data were obtained for 130 patients enrolled in the Enara Health weight loss programme during the study period, of whom 129 (99.2%) met the study inclusion criteria and had valid data. Of the 129 participants, 71 (55%) AOM users were included in the subgroup (completers) analyses.

Baseline patient characteristics

Table 1 presents the baseline characteristics of the total study cohort (ITT $N = 129$) and subgroups by AOM usage status. The mean age was 45 ± 14 years. The mean weight and mean BMI were 103 ± 23 kg and 37 ± 6 kg m^{-2}, respectively. Women comprised approximately two-thirds (67%) of the cohort. Compared with the non-users ($N = 58$) at baseline, AOM users on average were significantly heavier (107.0 ± 22.0 vs. 97.0 ± 20.0 kg) and featured a higher proportion of men (41% vs. 24%), increased BMI (38.0 ± 6.0 vs. 35.0 ± 5.0 kg m^{-2}) and lower high-density lipoprotein cholesterol level (51.0 ± 11.7 vs. 57.0 ± 14.4 mg dL^{-1}) (all differences, $P < 0.05$). Seventy-seven per cent of AOM users and 71% of non-users choose to do the intensive weight loss programme. The difference between the two cohorts was not statistically significant ($P = 0.356$).

Of the 71 AOM users, 50 (70.4%) had a weigh-in at 72 weeks and were considered completers (Table 2). Medication utilization among the AOM user subgroup was as follows: phentermine/TPM extended-release ($n = 44$, 62%), phentermine ($n = 17$, 24%), liraglutide ($n = 12$, 17%), bupropion/naltrexone ($n = 9$, 13%) and TPM ($n = 5$, 7%). Twenty-two per cent ($n = 16$) of AOM users used a combination of two AOMs. Of the 50 completers, 22 (44%) were early AOM starters and 28 (56%) were delayed AOM starters. With respect to treatment duration, 43 of the 50 completers (86%) had been treated for >6 months; seven (14%) had been treated for ≤6 months. While the baseline characteristics did not differ significantly between early and delayed AOM starters, completers who had been treated for a long duration had a significantly higher mean BMI compared with those treated for a short duration (39 ± 6.8 vs. 33 ± 3.2 kg m^{-2}; $P < 0.05$).

Primary endpoint analysis

Comparison of changes in baseline weight between anti-obesity medication users versus non-users at week 72 (intention-to-treat cohort)

At week 72, AOM users ($N = 71$) achieved significantly greater mean percentage and absolute reductions in baseline weight than non-users ($N = 58$) (Figures 2 and 3 and Table 3). On average, the baseline weight was decreased by 14.0 ± 6.2% in AOM users compared with 10.9 ± 6.8% in non-users ($P = 0.008$). The mean absolute weight loss was 15.0 ± 7.2 kg for AOM users and 10.4 ±

© 2019 The Authors
Obesity Science & Practice published by World Obesity and The Obesity Society and John Wiley & Sons Ltd Obesity Science & Practice

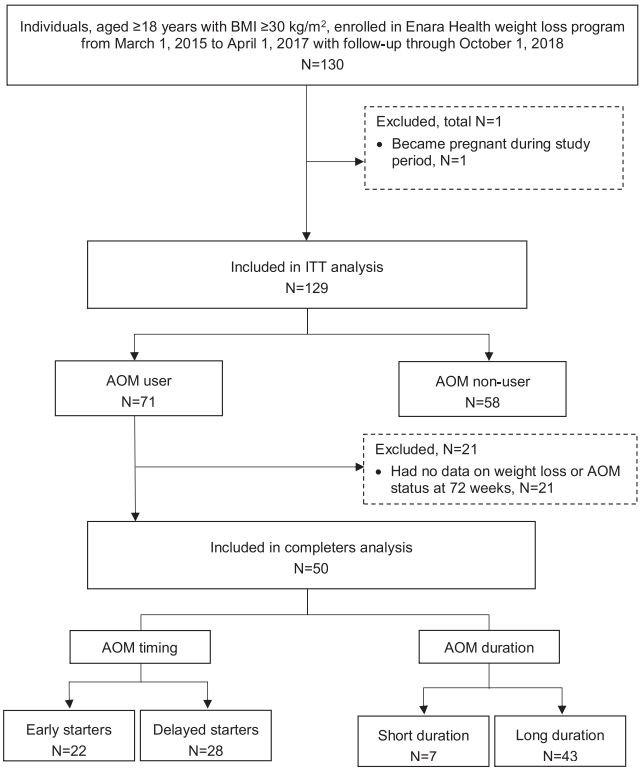


Figure 1 Study cohort selection flow diagram. AOM, anti-obesity medication; BMI, body mass index; ITT, intent-to-treat.

Table 1 Participants' baseline characteristics

	All users (*n* = 129)	Non-medication users (*n* = 58)	Medication users (*n* = 71)
Age (years)	45 ± 14	46 ± 14	44 ± 12
Baseline weight (kg)*	103 ± 23	97 ± 20	107 ± 22
BMI (kg m^{-2})*	37 ± 6	35 ± 5	38 ± 6
Gender (male), *n* (%)*	43 (33%)	14 (24%)	29 (41%)
Gender (female), *n* (%)*	86 (67%)	44 (76%)	42 (59%)
HbA1C (%)	5.70 ± 0.51 (*n* = 105)	5.63 ± 0.31 (*n* = 48)	5.75 ± 0.62 (*n* = 57)
Total cholesterol (mg dL^{-1})	200.38 ± 38.55 (*n* = 113)	205.52 ± 36.56 (*n* = 52)	196.0 ± 39.65 (*n* = 61)
Triglycerides (mg dL^{-1})	131.25 ± 63.92 (*n* = 114)	127.49 ± 56.38 (*n* = 53)	134.53 ± 69.67 (*n* = 61)
LDL cholesterol (mg dL^{-1})	120.09 ± 33.20 (*n* = 112)	123.58 ± 31.46 (*n* = 52)	117.07 ± 34.36 (*n* = 60)
HDL cholesterol (mg dL^{-1})*	53.71 ± 13.33 (*n* = 112)	56.98 ± 14.39 (*n* = 51)	50.98 ± 11.69 (*n* = 61)
Glucose (mg dL^{-1})	98.0 ± 14.49 (*n* = 100)	96.48 ± 9.73 (*n* = 43)	99.14 ± 17.15 (*n* = 57)

Values shown are *n* (%) or means ± standard deviation.
*Categories differ significantly from each other at $P < 0.05$.
BMI, body mass index; HbA1c, haemoglobin A1c; HDL, high-density lipoprotein; LDL, low-density lipoprotein.

© 2019 The Authors
Obesity Science & Practice published by World Obesity and The Obesity Society and John Wiley & Sons Ltd Obesity Science & Practice

Table 2 Medication users' baseline characteristics (completers only)

	All medication users (n = 50)	Early starters (n = 22)	Delayed starters (n = 28)	Short duration (n = 7)	Long duration (n = 43)
Age (years)	45 ± 13.04	45 ± 12.2	45 ± 13.66	44 ± 14.93	45 ± 12.69
Baseline weight (kg)	107 ± 24.9	114 ± 30.54	102 ± 17.49	92 ± 11.06	110 ± 25.68
BMI (kg m^{-2})	38 ± 6.7	39 ± 7.63*	37 ± 5.59*	33 ± 3.24	39 ± 6.78
Gender (male), n (%)	20 (40%)	6 (27%)	14 (50%)	4 (57%)	16 (37%)
Gender (female), n (%)	30 (60%)	16 (73%)	14 (50%)	3 (43%)	27 (63%)
HbA1C (%)	5.82 ± 0.41 (n = 40)	5.84 ± 0.48 (n = 19)	5.81 ± 0.36 (n = 21)	5.82 ± 0.33 (n = 6)	5.83 ± 0.43 (n = 34)
Total cholesterol (mg dL^{-1})	199.67 ± 37.04 (n = 43)	198.72 ± 34.17 (n = 18)	200.36 ± 38.96 (n = 25)	213.86 ± 56.23 (n = 7)	196.92 ± 31.26 (n = 36)
Triglycerides (mg dL^{-1})	142.82 ± 69.31 (n = 43)	154.44 ± 76.96 (n = 18)	134.46 ± 61.90 (n = 25)	143.86 ± 83.98 (n = 7)	142.62 ± 66.08 (n = 36)
LDL cholesterol (mg dL^{-1})	118.86 ± 31.95 (n = 42)	114.22 ± 27.52 (n = 18)	122.33 ± 34.50 (n = 24)	137.0 ± 49.91 (n = 7)	115.23 ± 25.46 (n = 35)
HDL cholesterol (mg dL^{-1})	51.46 ± 11.76 (n = 43)	54.18 ± 12.69 (n = 18)	49.52 ± 10.62 (n = 25)	45.29 ± 13.18 (n = 7)	52.67 ± 11.07 (n = 36)
Glucose (mg dL^{-1})	99.93 ± 16.00 (n = 41)	95.74 ± 12.63 (n = 19)	103.54 ± 17.65 (n = 22)	101.0 ± 6.07 (n = 5)	99.78 ± 16.92 (n = 36)

Values shown are n (%) or means ± standard deviation.
*Categories differ significantly from each other at $P < 0.05$.
BMI, body mass index; HbA1c, haemoglobin A1c; HDL, high-density lipoprotein; LDL, low-density lipoprotein.

6.9 kg for non-users ($P < 0.001$). Compared with non-users, a markedly higher proportion of AOM users attained weight reductions ≥10% of the baseline (69% vs. 50%; $P = 0.03$), as well as ≥15% of the baseline (45.1% vs. 19.0%, respectively; $P < 0.001$).

Subgroup comparisons of changes in baseline weight at week 72 (completers)

At week 72, AOM early starters ($N = 22$) achieved significantly greater mean percentage and absolute reductions in baseline weight than delayed starters ($N = 28$) (Figure 4 and Table 4). On average, the baseline weight was decreased by 17.6 ± 5.3% in early starters compared with 14.0 ± 5.5% in delayed starters ($P = 0.024$). The mean absolute weight loss was 19.7 ± 6.7 kg for early starters and 14.1 ± 5.6 kg for delayed starters ($P = 0.003$). A markedly higher proportion of early starters attained weight reductions ≥15% of the baseline compared with delayed starters (72.3% vs. 42.9%, respectively; $P = 0.03$).

The mean percentage reduction in baseline weight at 72 weeks was numerically greater for completers treated with AOM for a long duration ($N = 43$) than for those treated for a short duration ($N = 7$) (−16.0 ± 5.7% vs. −12.7 ± 4.2%, respectively), albeit not statistically different (Table 4). The mean absolute weight loss, however, was significantly increased with long treatment duration (17.3 ± 6.7 vs. 11.8 ± 4.2 kg; $P = 0.019$). The proportions of patients achieving categorical percentage losses did not differ significantly between the subgroups.

Analysis of secondary endpoints

Comparison of changes in baseline weight between anti-obesity medication users versus non-users at weeks 24 and 48 (intention-to-treat cohort)

At both weeks 24 and 48, AOM users achieved significantly greater mean percentage and mean absolute reductions in baseline weight compared with non-users. On average, baseline weight was reduced by 13.8% in AOM users and by 11.1% in non-users at week 24 ($P = 0.009$) and by 13.4% and 10.6% in the respective subgroups at week 48 ($P = 0.016$). At week 24, the mean absolute weight loss was 14.7 ± 6.9 kg for AOM users and 10.7 ± 5.9 kg for non-users ($P = 0.001$). At week 48, the mean absolute weight loss was 14.4 ± 7.3 kg for AOM users and 10.1 ± 6.4 kg for non-users ($P = 0.001$). Significantly higher proportions of AOM users lost ≥15% of their baseline weight than non-users at both week 24 (40.9% vs. 22.4%; $P = 0.03$) and week 48 (39.4% vs. 17.2%; $P = 0.01$).

© 2019 The Authors
Obesity Science & Practice published by World Obesity and The Obesity Society and John Wiley & Sons Ltd Obesity Science & Practice

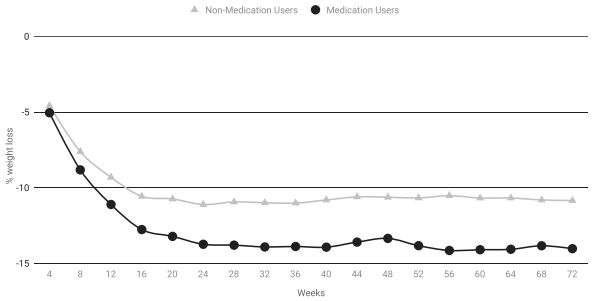


Figure 2 Differences in mean percentage weight reduction over 72 weeks between anti-obesity medication users and non-users. Estimated mean percentage reduction in baseline weight over 72 weeks in the intention-to-treat cohort (71 anti-obesity medication users and 58 non-users). *P*-values for differences between the subgroups (per *t*-test) at week 24, week 48 and week 72 are shown in Table 3.

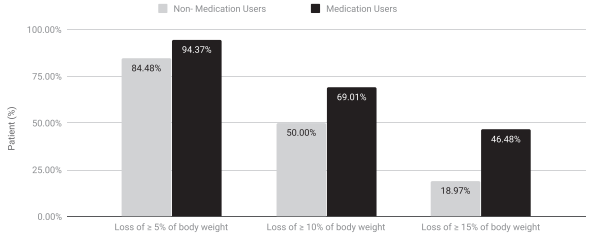


Figure 3 Categorical (5%, 10% and 15%) weight loss at week 72 in anti-obesity medication users versus non-users. Differences in the proportions of patients with 10% and 15% categorical weight loss between anti-obesity medication users and non-users were statistically significant ($P < 0.05$, per Chi-square tests) (Table 3).

Table 3 Changes in primary endpoints

	Non-medication users (n = 58)	Medication users (n = 71)	*P*-value
Change in weight (%)			
Week 24	−11.12 ± 5.15	−13.76 ± 6.05	0.009*
Week 48	−10.64 ± 6.13	−13.36 ± 6.32	0.016*
Week 72	−10.86 ± 6.84	−14.04 ± 6.22	0.008*
Change in weight (kg)			
Week 24	−10.72 ± 5.90	−14.74 ± 6.94	0.001*
Week 48	−10.13 ± 6.35	−14.43 ± 7.32	0.001*
Week 72	−10.35 ± 6.90	−14.99 ± 7.19	<0.001*
Loss of ≥5% of body weight (%)			
Week 24	93.10%	95.77%	0.51
Week 48	84.48%	92.96%	0.12
Week 72	84.48%	94.29%	0.06

Continues

© 2019 The Authors
Obesity Science & Practice published by World Obesity and The Obesity Society and John Wiley & Sons Ltd Obesity Science & Practice

Table 3. *Continued*

	Non-medication users (*n* = 58)	Medication users (*n* = 71)	*P*-value
Loss of ≥10% of body weight (%)			
Week 24	53.45%	69.01%	0.07
Week 48	51.72%	59.15%	0.31
Week 72	50.00%	69.01%	0.03*
Loss of ≥15% of body weight (%)			
Week 24	22.41%	40.85%	0.03*
Week 48	17.24%	39.44%	0.01*
Week 72	18.97%	45.07%	<0.001*

Values shown are *n* (%) or means ± standard deviation for the intention-to-treat population (*N* = 129).
*Categories differ significantly from each other at $P < 0.05$.

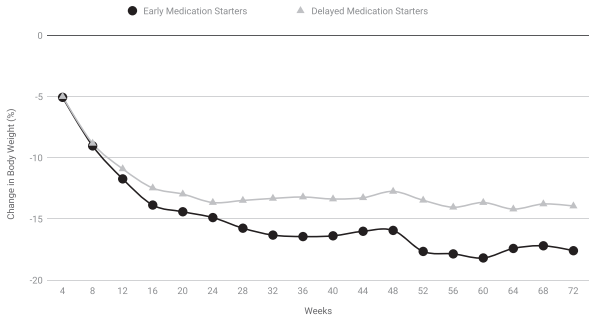


Figure 4 Differences in mean per cent weight reduction over 72 weeks between anti-obesity medication early starters versus delayed starters. Estimated mean percentage reduction in baseline weight over 72 weeks in anti-obesity medication completers (*N* = 50) with early start (*N* = 22) and delayed start (*N* = 28) of treatment. *P*-values (per *t*-tests) for differences between the subgroups at week 24, week 48 and week 72 are shown in Table 4.

Table 4 Changes in primary endpoints among medication users

	Early starters (*n* = 22)	Delayed starters (*n* = 28)	*P*-value	Short duration (*n* = 7)	Long duration (*n* = 43)	*P*-value
Change in weight (%)						
Week 24	−14.90 ± 6.22	−13.68 ± 5.73	0.425	−14.98 ± 6.03	−14.15 ± 6.01	0.688
Week 48	−15.95 ± 6.52	−12.76 ± 4.58	0.047*	−13.39 ± 3.75	−14.40 ± 6.09	0.536
Week 72*	−17.6 ± 5.30	−13.95 ± 5.48	0.024*	−12.67 ± 4.21	−16.03 ± 5.77	0.114
Change in weight (kg)						
Week 24	−16.12 ± 6.75	−14.49 ± 7.15	0.361	−15.58 ± 7.47	−15.27 ± 6.88	0.9
Week 48	−17.39 ± 7.12	−13.22 ± 5.45	0.021*	−13.29 ± 5.10	−15.52 ± 6.81	0.298
Week 72*	−19.67 ± 6.66	−14.07 ± 5.59	0.003*	−11.82 ± 4.16	−17.30 ± 6.71	0.019*
Loss of ≥5% of body weight (%)						
Week 24	96.97%	96.77%	0.96	91.67%	98.08%	0.25
Week 48	100.00%	96.67%	0.35	100.00%	97.87%	0.66
Week 72*	100.00%	96.43%	0.373	100.00%	97.67%	0.68
Loss of ≥10% of body weight (%)						
Week 24	75.76%	67.74%	0.48	75.00%	71.15%	0.79
Week 48	80.77%	56.67%	0.05*	77.78%	65.96%	0.49
Week 72*	90.91%	75.00%	0.15	71.43%	83.72%	0.43

Continues

© 2019 The Authors
Obesity Science & Practice published by World Obesity and The Obesity Society and John Wiley & Sons Ltd Obesity Science & Practice

Table 4. *Continued*

	Early starters (n = 22)	Delayed starters (n = 28)	P-value	Short duration (n = 7)	Long duration (n = 43)	P-value
Loss of ≥15% of body weight (%)						
Week 24	48.48%	41.94%	0.6	58.33%	42.31%	0.31
Week 48	53.85%	33.33%	0.12	44.44%	42.55%	0.92
Week 72*	72.73%	42.86%	0.03*	28.57%	58.14%	0.15

Values shown are n (%) or means ± standard deviation for completers analysis (N = 50) in the medication subgroup.
*Categories differ significantly from each other at $P < 0.05$.

Subgroup comparisons of changes in baseline weight at weeks 24 and 48 (completers)

At week 24, early starters (N = 22) exhibited numerically higher mean percentage and absolute reductions in baseline weight compared with delayed starters (N = 28), albeit the differences were not significant (Table 4). The proportions of patients achieving categorical percentage losses did not differ significantly between the subgroups. At week 48, however, early starters achieved marked percentage and absolute reductions in baseline weight compared with delayed starters. On average, the baseline weight was decreased by 16% in early starters compared with 12.8% in delayed starters (P = 0.047). The mean absolute weight loss was 17.4 ± 7.1 kg for early starters and 13.2 ± 5.5 kg for delayed starters (P = 0.021). A higher proportion of early starters attained weight reductions ≥10% of the baseline at week 48 compared with delayed starters (80.8% vs. 56.7%, respectively; P = 0.05).

At both weeks 24 and 48, the AOM subgroups categorized by treatment duration did not differ significantly with respect to the magnitude of changes in baseline weight or the proportions achieving categorical percentages of weight loss (Table 4).

Discussion

This study supports that AOM has an incremental impact on weight loss efforts with lifestyle modification (17,18). The addition of AOM by patients was associated with a greater mean percentage reduction in weight (14.2% vs. 10.8%) and a higher proportion achieving 15% weight loss (45.1% vs. 19.0%). This was not a randomized clinical trial and was not restricted to one medication; rather, it is a study of the real-world application of obesity medications where a clinician works to find the optimal medications for a patient based on past medical history, contraindications, coverage and preference. In a handful of studies, higher initial BMI is associated with less weight loss after surgical interventions (19,20). However, the impact of initial BMI on AOM efficacy has not been studied. In our study, the association of AOM with greater per cent weight loss in patients who were heavier at baseline (and had greater disease burden) suggests that BMI-adjusted differences are likely to be even greater.

More importantly, this study adds to the growing body of indirect evidence elucidated in the introduction showing that medication timing and duration may perhaps play a significant role in outcomes. Patient who began an AOM at the onset of their behavioural modification achieved an impressive 17.6% weight loss at 72 weeks; 72% of this subset achieved >15% weight loss. We believe that patients who begin an AOM at onset and continue to engage in behaviour modification may experience less metabolic adaptation as they lose weight (6,21–23). A prospective study examining early versus delayed medication initiation on metabolic adaptation parameters would be very informative. If newer obesity medications are found to play a role in temporizing the onset of metabolic adaptation, then earlier introduction during a patient's weight loss journey may help sustain outcomes.

Our findings differ from studies demonstrating additional weight loss when TPM or liraglutide were added after low-calorie-induced weight loss (7,9) compared with studies initiating therapy at programme onset (8,10). The difference in weight loss observed may likely be explained by varying dietary and behavioural protocols rather than delayed onset of medication. Alternatively, in real-world application, early AOM initiation may reinforce dietary and behavioural plans and encourages patients to engage longer.

While medication duration was not found to be significant, greater weight loss and categorical weight loss were observed in subgroup receiving treatment for a long duration. Additionally, the short duration subgroup displayed some evidence of weight recidivism at week 72. This correlates with studies showing that when AOM is stopped, weight increases gradually to the level seen with lifestyle changes alone (24).

The level of weight loss observed in this study is greater than that observed in AOM randomized controlled trials; however, the difference between AOM user and non-users subgroups observed in the present study was similar to that reported in the literature (17,18,25). The varying weight loss outcomes between studies is an added indication that dietary, behavioural and

© 2019 The Authors
Obesity Science & Practice published by World Obesity and The Obesity Society and John Wiley & Sons Ltd Obesity Science & Practice

educational components of lifestyle modification interventions may enhance outcomes.

Study strengths include a 72-week treatment duration, which allowed for exploration of the downstream impact of AOMs, and the high completer rate (71%) for patients electing treatment with AOM. Potential limitations include the retrospective nature, small sample sizes in the completer subgroups and the gender and BMI differences between AOM users and non-users in the ITT analysis. Differences in starting BMI and weight between AOM user and non-users were adjusted by using per cent weight loss rather than absolute weight loss as the primary endpoint. These differences are also informatory as it suggests that in real-world settings, male patients and patients who have greater obesity burden are more open to using AOM.

Precision medicine calls for the use of genomic and metabolic data to personalize medication delivery (26). In the context of obesity, the timing and duration of medication against the backdrop of social, behavioural, nutritional and environmental factors become additional precision elements (26). Obesity is complicated disease that will require multimodal lifestyle modification interventions, and our understanding of why some individuals respond well to certain treatments while others do not is limited (27). While current focus is on how genomic, behavioural, biological, environmental and psychosocial factors impact treatment, we must also consider how different treatment(s) interact and impact one another. The different combinations and order in which therapies are delivered present an opportunity to optimize outcomes.

Acknowledgements

Erin Williams of Marinus Consulting, LLC assisted with the editing of this manuscript. Enara Health provided no funding for her work.

Funding

This work was funded by Enara Health.

Conflict of Interest Statement

Rojin Safavi was not compensated for this study nor does she have any conflicts of interest. The remaining authors all work for Enara Health and do not have any other declarations.

References

1. Stegenga H, Haines A, Jones K, Wilding J. Identification, assessment, and management of overweight and obesity: summary of updated NICE guidance. *BMJ* 2014; **349**: g6608.
2. Jensen MD, Ryan DH, Apovian CM, et al. 2013 AHA/ACC/TOS guideline for the management of overweight and obesity in adults: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines and The Obesity Society. *J. Am. Coll. Cardiol.* 2014; **63**: 2985–3023.
3. Apovian CM, Aronne LJ, Bessesen DH, et al. Pharmacological management of obesity: an Endocrine Society clinical practice guideline. *J. Clin. Endocrinol. Metab.* 2015; **100**: 342–362.
4. American Diabetes Association. 7. Obesity management for the treatment of type 2 diabetes: standards of medical care in diabetes—2018. *Diabetes Care* ; **41**: S65–S72.
5. Curry SJ, Krist AH, Owens DK, et al. Behavioral weight loss interventions to prevent obesity-related morbidity and mortality in adults: US preventive services task force recommendation statement. *JAMA* ; **320**: 1163–1171.
6. Bessesen DH, Van Gaal LF. Progress and challenges in anti-obesity pharmacotherapy. *Lancet Diabetes Endocrinol.* 2018; **6**: 237–248.
7. Boutari C, Polyzos SA, Mantzoros CS. Of mice and men: why progress in the pharmacological management of obesity is slower than anticipated and what could be done about it? *Metabolism* ; **96**: vi–xi.
8. Wadden TA, Walsh OA, Berkowitz RI, et al. Intensive behavioral therapy for obesity combined with liraglutide 3.0 mg: a randomized controlled trial. *Obesity (Silver Spring)* 2019; **27**: 75–86.
9. Wadden TA, Berkowitz RI, Sarwer DB, Prus-Wisniewski R, Steinberg C. Benefits of lifestyle modification in the pharmacologic treatment of obesity: a randomized trial. *Arch. Intern. Med.* 2001; **161**: 218–227.
10. Astrup A, Caterson I, Zelissen P, et al. Topiramate: long-term maintenance of weight loss induced by a low-calorie diet in obese subjects. *Obes. Res.* 2004; **12**: 1658–1669.
11. Kramer CK, Leitao CB, Pinto LC, Canani LH, Azevedo MJ, Gross JL. Efficacy and safety of topiramate on weight loss: a meta-analysis of randomized controlled trials. *Obes. Rev.* 2011; **12**: e338–e347.
12. Wadden TA, Hollander P, Klein S, et al. Weight maintenance and additional weight loss with liraglutide after low-calorie-diet-induced weight loss: the SCALE Maintenance randomized study. *Int. J. Obes. (Lond)* 2013; **37**: 1443–1451.
13. Pi-Sunyer X, Astrup A, Fujioka K, et al. A randomized, controlled trial of 3.0 mg of liraglutide in weight management. *N. Engl. J. Med.* 2015; **373**: 11–22.
14. Sharma VD, Combes FP, Vakilynejad M, Lahu G, Lesko LJ, Trame MN. Model-based approach to predict adherence to protocol during antiobesity trials. *J. Clin. Pharmacol.* 2018; **58**: 240–253.
15. Garvey WT, Ryan DH, Look M, et al. Two-year sustained weight loss and metabolic benefits with controlled-release phentermine/topiramate in obese and overweight adults (SEQUEL): a randomized, placebo-controlled, phase 3 extension study. *Am. J. Clin. Nutr.* 2012; **95**: 297–308.
16. Lewis KH, Fischer H, Ard J, et al. Safety and effectiveness of longer-term phentermine use: clinical outcomes from an electronic health record cohort. *Obesity (Silver Spring)* 2019; **27**: 591–602.
17. Wadden TA, Berkowitz RI, Womble LG, et al. Randomized trial of lifestyle modification and pharmacotherapy for obesity. *N. Engl. J. Med.* 2005; **353**: 2111–2120.
18. Khera R, Murad MH, Chandar AK, et al. Association of pharmacological treatments for obesity with weight loss and adverse events: a systematic review and meta-analysis. *JAMA* 2016; **315**: 2424–2434.

© 2019 The Authors
Obesity Science & Practice published by World Obesity and The Obesity Society and John Wiley & Sons Ltd Obesity Science & Practice

Chapter 3

Mobile Engagement Behaviors Associated with Sustained 15Weight Loss

The text of this chapter includes reprints of the previously published material. Lih A, Safavi R, et al. (2020). My contribution to this study involves statistical analysis of contributing weight loss factors in patients participating in Enara mobile application (mHealth).

Journal of Obesity and Chronic Diseases

https://doi.org/10.17756/jocd.2020-035

Research Article

Open Access

Mobile Engagement Behaviors Associated with Sustained 15% Weight Loss

Ann Lih[1], Rojin Safavi[2], Zariyah Isaq[1], Shery Haller[1] and Mohammed Bailony[1*]

[1]*Enara Health, 1650 S. Amphlett Blvd, Suite 115, San Mateo, CA 94402, USA*

[2]*University of California, Santa Cruz, California, USA*

***Correspondence to:**
Mohammed Rami Bailony, M.D., M.S
CEO and Co-Founder, Enara Health
1650 S. Amphlett Blvd, Suite 115
San Mateo, CA 94402, USA
Tel: +1-650-319-8654
E-mail: rami@enarahealth.com

Received: February 01, 2020
Accepted: May 01, 2020
Published: May 04, 2020

Citation: Lih A, Safavi R, Isaq Z, Haller S, Bailony M. 2020. Mobile Engagement Behaviors Associated with Sustained 15% Weight Loss. *J Obes Chronic Dis* 4(1): 35-41.

Copyright: © 2020 Lih et al. This is an Open Access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY) (http://creativecommons.org/licenses/by/4.0/) which permits commercial use, including reproduction, adaptation, and distribution of the article provided the original author and source are credited.

Published by United Scientific Group

Abstract

Background: This study aimed to report weight loss outcomes and evaluate behaviors of high performing participants who achieved ≥ 15% weight loss at 6 and 18 months in a precision medicine weight loss program, which incorporates clinical visits, a mobile application (mHealth) and tailored nutritional programming.

Methods: A retrospective analysis was performed on participants, aged ≥ 18 years with BMI ≥ 25 kg/m^2, who enrolled in Enara Health's weight loss program from March 1, 2015 to October 1, 2018. Participants were matched into 6 and 18-month cohorts. An analysis characterized high-performing participants who lost ≥ 15% of their starting weight. Behaviors evaluated included: monthly appointments, monthly coaching text conversations, weekly food pictures, percent of educational content completed, weekly weight monitoring, and number of days wearing an activity tracker.

Results: Average weight loss for the 6-month (n = 479) and 18-month (n = 145) cohorts was -12.40% and -13.99%, respectively. 32.57% of participants achieved 15% weight loss at 6 months and 41.38% of participants achieved 15% weight loss at 18 months. High achievers at 6 months had increased monthly appointments (3.05 ± 0.82 *vs* 2.72 ± 0.92, P < 0.001), weekly food pictures (40.64 ± 27.85 *vs* 33.61 ± 26.51, P = 0.011), educational activity compliance (60.98 ± 34.48% *vs* 47.10 ± 34.91%, P < 0.001) and weekly weights (3.11 ± 2.28 *vs* 2.55 ± 1.97, P = 0.011). High achievers at 18 months attended more monthly appointments (2.12 ± 0.58 *vs* 1.78 ± 0.70, P = 0.002), engaged in more monthly coaching text conversations (6.92 ± 5.61 *vs* 4.41 ± 3.65, P = 0.002), and weighed themselves more frequently beyond 12 months (1.92 ± 1.98 *vs* 1.26 ± 1.64, P = 0.47).

Conclusion: Participants in the Enara Health precision weight loss program lost an average of- 13.99% at 18 months. Self-tracking behaviors characterized high achievers at 6 months and increased number of provider-patient touchpoints characterized high achievers at 18 months. Mobile health applications may be better suited to achieve sustained weight loss by designing features that facilitate provider-patient interactions on top of self-tracking technology.

Keywords

mHealth, Mobile health, Obesity treatment, Weight loss, Behavioural modifications

Introduction

Recent studies have predicted that by 2030 nearly 1 in 2 Americans will struggle with obesity [1]. The exponential growth in incidence rates is correlated to

an increasing chronic disease burden and decreasing mortality after the age of 40 in this population [2]. Problematically what is often underestimated with respect to this growing epidemic is the difficulty of treatment.

Standalone approaches to obesity, whether economical, behavioral, lifestyle or pharmacological have not been able to demonstrate sustained outcomes. A recurring trend throughout these weight-loss interventions was poor attrition past 6-months [3, 4]. Several studies have shown that both self-tracking technology and programs with multicomponent approaches help reduce attrition in weight loss programs [5, 6]. The opportunity for mobile technology to help deliver and tailor multimodal obesity treatments has been widely explored over the last decade [7-9]. Almost 30% of all mobile health development has focused exclusively or partially on weight loss. We now have approximately 30,000 mobile applications to help with weight loss [10].

Research evaluating the effectiveness of mobile health interventions have demonstrated conflicting findings on whether technology improves weight loss outcomes [11-13]. Some meta-analysis of randomized control trials has found that web-based interventions did not improve weight loss over off-line counterparts [14] while others have found that they do exhibit a modest improvement in results [15-17]. The reason for this conflicting data is largely because mobile interventions are lumped monolithically when in fact they are widely divergent in the type of intervention they each deliver. Each application has varying degrees of self-monitoring, feedback, gamification, social support, and educational features. Current mobile health studies on weight loss have not taken this into consideration and as a result we do not have a good understanding on which features are most effective and in which settings [18, 19] Without a more nuanced examination, decision makers and developers lack information on how to improve mobile weight loss interventions. One study that did examine the impact of varying intervention features within a mobile application found text message based coaching conversation was associated with increased weight loss in females but not in males. This study is the first mobile health (mHealth) study that postulates that a mobile health intervention may need to be personalized by gender [20].

Enara Health is a comprehensive program, delivering behavioral, pharmacologic, and nutritional therapy via mobile and web-based application. The purpose of this study was to evaluate mobile engagement behaviors of high performing participants who achieved ≥ 15% weight loss at 6 and 18 months in a precision medicine weight loss program. The mobile application under study allows for examination and comparison of self-monitoring, feedback, support, and educational features within the same mHealth program. This may shed insight on which features and usage are better for sustained weight loss.

Methods

Study design

This is a retrospective cohort study analysing the mobile engagement behaviors of adults participating in a comprehensive precision weight loss program (Enara Health). Data were collected from the electronic health records of patients who enrolled and started a weight loss journey during the period from 1 March 2015 to 1 October 2018 (with follow-up through 1 April 2019). Enara Health provided the study investigators with all relevant data following de-identification of the entire dataset. The Hummingbird Institutional Review Board (Needham, MA, USA) approved the study protocol and deemed that the study met the requirements for exemption from the US Department of Health and Human Services regulations for the protection of human subjects (45 CFR 46.101[b][4]).

Weight loss program interventions

Enara Health is a medical company based in San Mateo, California, that offers a unique hybrid digital and in-person weight loss program. The Enara Health weight loss program has been operating since March 2015. By combining mobile technology with healthcare visits, the program provides patients with personalized physician-driven medical weight loss plans. Upon enrolment, patients are initially evaluated by a physician or physician assistant who determines the severity of obesity and the patient's readiness to change and performs a full metabolic workup and examination. Patients engage in one-on-one meetings with a registered dietitian on a weekly basis during the first 3 months, once every other week during the next 3 months and then once monthly for 6 months. Patients also see their physician or physician assistant on average once a month for the first 3 months, and then every 3 months thereafter. Patients may conduct their visits either in-person or via face-to-face video communication. There is a monthly fee of $50 for patients with insurance (most insurances accepted) or $300 without insurance.

Between visits, patients utilize a mHealth application, which delivers tailored educational content, messages of encouragement and support, nutritional and behavioral feedback, a review of meals and feedback by a registered dietitian, as well as exercise suggestions, planning and encouragement by a fitness trainer. Messages can be initiated by either patient or providers (physicians, dietitians, fitness trainers). Providers will typically respond within a 12-24-hour window; while it is not a 24-hour on-call messaging service, it allows for sufficient accountability, guidance, and support to patients throughout their week. Providers can also initiate conversations in response key alerts they receive on patients' appointment, exercise, food log, and weight trends. Patients may choose between an intense, rapid weight loss program (daily caloric intake: 800-1,000 kcal) or a non-intensive weight loss program that emphasizes non-processed foods but does not encourage calorie counting. Both options encourage whole, unprocessed, low glycemic foods. Although all participants are assigned a target exercise goal of 150 min week-1, a personal exercise coach determines the duration of exercise based on the patient's physical ability. Many patients utilize activity trackers such as Fitbit® or Apple Watch® to monitor weekly exercise.

Where appropriate, patients are prescribed medications

to assist with weight loss. Prescriptions are driven by the patient's request after attending an educational visit on anti-obesity medications (AOMs). AOMs are prescribed in accordance with the US Food and Drug Administration-labeled recommendations; however, in the event of non-coverage, generic equivalents are prescribed, and any off-label usage is discussed with the patient. The choice of medication is ultimately made by the patients and their physicians.

Prescribed weight loss medications include lorcaserin, phendimetrazine, diethylpropion, phentermine/topiramate, liraglutide and bupropion/naltrexone. Some patients may be prescribed more than one AOM. Routine follow-up visits with the provider are conducted to ensure that the patient is responding positively to the medication and is not experiencing adverse effects. All patients are started on the lowest dosage of medications, while dose adjustments are made on an individual basis as needed to achieve the desired outcomes.

Study cohort selection

Patients enrolled in the Enara Health weight loss program during the study period, who were aged ≥ 18 years with a BMI ≥ 25 kg/m^2 at enrollment, were eligible for inclusion. Patients who became pregnant during the study period were excluded from the analyses.

Assignment of mobile engagement cohort

Overall, participants were matched into a 6 and 18 month cohorts so long as their start date indicated that they could have participated in the study duration regardless of their engagement status at that time point. Completers were defined as any participant who logged a home weight ± 30 days from the cohort end date. High achievers were defined as participants who achieved greater than 15% weight loss.

Study data collection

Enara Health collected demographic and clinical data for all study patients at the initial clinic visit. Height, body weight, blood pressure, pulse and body composition were measured at the first appointment to establish baseline values and then rechecked in the clinic at every in-person appointment thereafter. Body weight was collected daily via the patient's home scale, which connected directly to Enara Health's provider dashboard. Mobile engagement data were collected from Enara's mobile app and included monthly appointments, monthly coaching text conversations, weekly food pictures, percent of educational content completed, weekly weight monitoring, and number of days wearing an activity tracker.

Study endpoints

The primary outcome measurements were total percent of weight lost and percentage of participants who were high achievers defined as those who lost ≥ 15% of their starting weight.

For each mobile engagement behavior in a subgroup, the average (± standard deviation) of the engagement values were calculated. The significance of the average engagement between the subgroups were measured using a t-test. All statistical analyses were performed using Python 3.6. Statistical significance was defined as $P < 0.05$.

For the 18-month cohort, sustained mobile engagement behavior was defined as behavior averages from months 12-18 within each subgroup. This allowed for exploration and comparison of behavioral markers of success at different time periods throughout a participant's journey.

Results

Selection of study cohort

Figure 1 shows the distribution and selection of the final study cohorts and high achiever subgroups. Data were obtained for 569 patients enrolled in the Enara Health weight loss program who met the study criteria during the study period. 84% (n = 479) of the 6-month cohort (n = 569) had weight data at the 6-month mark while 56% (n = 145) of the 18-month cohort (n = 261) had weight data at the 18-month mark. 32.57% (n = 156) of participants in the 6-month cohort were high achievers (≥ 15% weight loss) and 41.38% (n = 60) of participants in the 18-month cohort were high achievers (≥ 15% weight loss at 18 months).

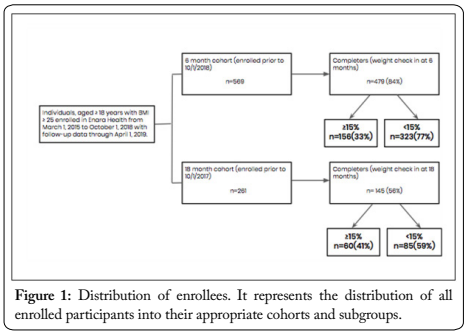


Figure 1: Distribution of enrollees. It represents the distribution of all enrolled participants into their appropriate cohorts and subgroups.

Baseline patient characteristics

Table 1 and 2 present the baseline characteristics between the subgroups of the 6-month cohort and 18-month cohort respectively. The mean age was 46 ± 13 years for both cohorts. In the 6-month cohort, high achievers had significantly higher starting BMI (35.90 ± 6.17 *vs* 34.24 ± 6.30, p = 0.007) and high triglycerides (143.50 ± 79.51 *vs* 121.34 ± 94.06, p = 0.018). There was no statistical difference between subgroups in the 18-month cohort for any baseline characteristics including anti-obesity medication use which was available for comparison in that cohort.

Comparison of changes in baseline weight in the 6-month cohort and 18-month cohort

Figure 2 and 3 demonstrate average weight loss for the 6-month (n = 479) and 18-month (n = 145) cohorts of -12.40% and -13.99% respectively. Figure 4 shows the percentages of the 6-month cohort and 18-month cohort that achieved 10%, 15% and 20% weight loss. 64.7% and 32.6% of participants

achieved a 10% and 15% weight loss at 6 months. 72% and 41.40% of participants achieved a 10% and 15% weight loss at 18 months.

Table 1: 6-month cohort participant's characteristics at baseline.

Behaviors	≥ 15% weight loss (n = 156)	< 15% weight loss (n = 323)	P value
BMI*	35.90 ± 6.17	34.24 ± 6.30	0.007*
Age (year)	46.22 ± 13.51	45.84 ± 12.87	0.771
Gender (male), (%)	29.48 (n = 46)	25.07 (n = 81)	-
Gender (female), (%)	70.51 (n = 110)	74.92 (n = 242)	-
HbA1C (%)	5.53 ± 1.17 (n = 120)	5.382 ± 1.67 (n = 221)	0.337
Total cholesterol (mg/dL)	196.48 ± 43.21 (n = 126)	189.36 ± 49.60 (n = 247)	0.155
Triglycerides (mg/dL)*	143.50 ± 79.51 (n = 125)	121.34 ± 94.06 (n = 247)	0.018*
LDL-cholesterol (mg/dL)	118.53 ± 34.51 (n = 123)	114.68 ± 38.60 (n = 245)	0.335
HDL-cholesterol (mg/dL)	50.71 ± 16.13 (n = 126)	52.09 ± 17.10 (n = 247)	0.449
Glucose (mg/dL)	95.28 ± 30.81 (n = 120)	97.34 ± 33.62 (n = 227)	0.568

Values shown are n (%) or means ± standard deviation. * categories differ significantly from each other at P < 0.05. BMI: Body Mass Index, HbA1c: Haemoglobin A1c, HDL: High-Density Lipoprotein, LDL: Low-Density Lipoprotein.

Table 2: 18-month cohort participant's characteristics at baseline.

Behaviors	≥ 15% weight loss (n = 60)	< 15% weight loss (n = 85)	P value
BMI	36.10 ± 5.01	34.69 ± 7.63	0.183
Medication Use (%)	0.78 ± 0.12 (n = 36)	0.66 ± 0.09 (n = 35)	0.291
Age (yr.)	44.23 ± 12.76	47.94 ± 13.37	0.095
Gender (male), (%)	26.66 (n = 16)	36.47 (n = 31)	-
Gender (female), (%)	73.33 (n = 44)	63.52 (n = 54)	-
HbA1C (%)	5.88 ± 1.26 (n = 46)	5.80 ± 0.99 (n = 65)	0.726
Total cholesterol (mg/dL)	207.38 ± 41.11 (n = 48)	197.30 ± 35.97 (n = 77)	0.170
Triglycerides (mg/dL)	150.17 ± 86.39 (n = 48)	145.60 ± 124.80 (n = 76)	0.812
LDL-cholesterol (mg/dL)	123.53 ± 35.71 (n = 47)	120.20 ± 34.94 (n = 77)	0.615
HDL-cholesterol (mg/dL)	54.42 ± 16.29 (n = 48)	54.10 ± 15.34 (n = 78)	0.915
Glucose (mg/dL)	102.09 ± 44.44 (n = 47)	104.21 ± 29.13 (n = 70)	0.775

Values shown are n (%) or means ± standard deviation. * categories differ significantly from each other at P < 0.05. BMI: Body Mass Index, HbA1c: Haemoglobin A1c, HDL: High-Density Lipoprotein, LDL: Low-Density Lipoprotein.

Engagement characteristics of high achievers at 6-months

High achievers at 6-months attended 11% more appointments (3.05 ± 0.82 *vs* 2.72 ± 0.92, P < 0.001), took 21% more food pictures (40.64 ± 27.85 *vs* 33.61 ± 26.51, p = 0.011), completed 29% more educational activities (60.98 ± 34.48 *vs* 47.10 ± 34.91, P < 0.001) and stepped on the scale 25% more times per week (3.11 ± 2.28 *vs* 2.55 ± 1.97, p = 0.011) compared to non-high achievers (Table 3). There were no differences in monthly provider text message conversations or the frequency of activity tracking between the two groups.

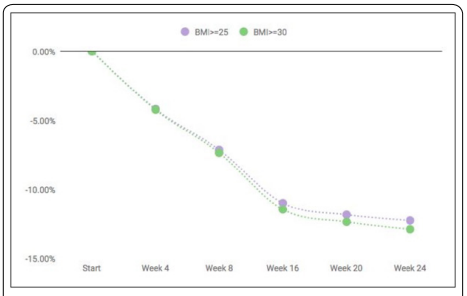


Figure 2: Average weight loss of 6-month cohort. It shows the average weight loss between the two subgroups of the 6-month cohort in 4-week intervals. The average weight loss for the entire 6-month cohort was-12.40% from baseline weight by week 24.

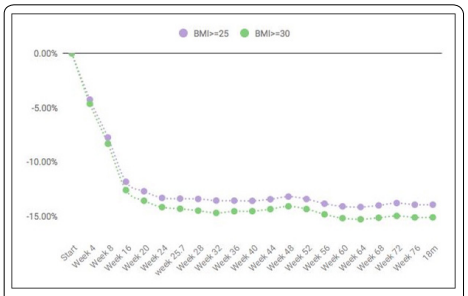


Figure 3: Average weight loss of 18-month cohort. It shows the average weight loss between the two subgroups of the 18-month cohort in 4-week intervals. The average weight loss for the entire 18-month cohort was-13.99% from baseline weight by 18 months.

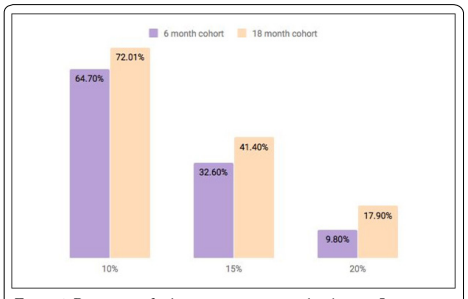


Figure 4: Percentage of cohort participants in each subgroup. It represents the percentage of cohort participants in each weight-loss subgroup.

Table 3: 6-month cohort-engagement behaviors of subgroups.

Behaviors	≥ 15% weight loss (n = 156)	< 15% weight loss (n = 323)	P value
Number of monthly appointments*	3.05 ± 0.82 (n = 156)	2.72 ± 0.92 (n = 323)	P < 0.001*
Number of monthly providers text conversations	10.37 ± 8.85 (n = 153)	9.76 ± 8.44 (n = 310)	0.486
Number of weekly foods pictures*	40.64 ± 27.85 (n = 150)	33.61 ± 26.51 (n = 306)	0.011*
Percent (%) of educational content completed*	60.98 ± 34.48 (n = 146)	47.10 ± 34.91 (n = 299)	P < 0.001*
Number of weekly weigh-ins*	3.11 ± 2.28 (n = 150)	2.55 ± 1.97 (n = 305)	0.011*
Number of days wearing an activity tracker	85.93 ± 56.36 (n = 124)	97.93 ± 56.89 (n = 128)	0.095

Values shown are n (%) or means ± standard deviation for the intention-to-treat population (N = 479). * categories differ significantly from each other at P < 0.05.

Engagement characteristics of high achievers at 18-months

High achievers at 18 months attended 19% more appointments (2.12 ± 0.58 *vs* 1.78 ± 0.70, P = 0.002) and initiated 57% more text-based conversations per month (6.92 ± 5.61 *vs* 4.41 ± 3.65) compared to non-high achievers. There was no statistical significance in food logging, exercise tracking, educational activities, and weight tracking between the two groups (Table 4).

Table 4: 18-month cohort-engagement behaviors of subgroups.

Behaviors	≥ 15% weight loss (n = 60)	< 15% weight loss (n = 85)	P value
Number of monthly appointments*	2.12 ± 0.58 (n = 60)	1.78 ± 0.70 (n = 85)	0.002*
Number of monthly provider text conversations*	6.92 ± 5.61 (n = 60)	4.41 ± 3.65 (n = 82)	0.003*
Number of weekly food pictures	25.56 ± 18.57 (n = 59)	21.12 ± 14.79 (n = 79)	0.136
Educational content completed (%)	53.21 ± 34.08 (n = 59)	50.87 ± 32.25 (n = 78)	0.687
Number of weekly weigh-ins	1.77 ± 1.79 (n = 60)	1.22 ± 1.61 (n = 77)	0.064
Number of days wearing an activity tracker	156.89 ± 147.58 (n = 75)	199.23 ± 174.15 (n = 43)	0.188

Values shown are n (%) or means ± standard deviation for the intention-to-treat population (N = 145). * categories differ significantly from each other at P < 0.05.

Sustained engagement characteristics high achievers at 18 months

A sub-analysis of engagement behaviors specifically between months 12 and months 18 showed the high achievers in the 18-month cohort were associated with 52% more weekly self-weigh-ins (1.92 ± 1.98 *vs* 1.26 ± 1.64, p = 0.047) and participated in 54% more text-based conversations with their providers (6.81 ± 5.90 *vs* 4.43 ± 3.8, p = 0.009) than non-high achievers (Table 5). When comparing activity from month 12-18 to months 1-12, weekly weigh-ins, educational activity, exercise tracking, monthly appointments, and food logging all dropped substantially while provider-patient text conversations persisted.

Table 5: 18-month cohort-sustained engagement characteristics.

Behaviors	≥ 15% weight loss (n = 60)	< 15% weight loss (n = 85)	P value
Monthly appointments	0.58 ± 0.29 (n = 60)	0.51 ± 0.29 (n = 77)	0.114
Monthly provider text conversations*	6.81 ± 5.90 (n = 60)	4.43 ± 3.83 (n = 74)	0.009*
Weekly food pictures	15.62 ± 18.37 (n = 59)	11.48 ± 14.23 (n = 71)	0.164
Percent of educational content completed	38.47 ± 37.22 (n = 59)	35.88 ± 38.74 (n = 70)	0.702
Weekly weight monitoring*	1.92 ± 1.98 (n = 60)	1.26 ± 1.64 (n = 69)	0.047*
Number of days wearing an activity tracker	65.56 ± 114.93 (n = 52)	91.13 ± 121.84 (n = 47)	0.292

Values shown are n (%) or means ± standard deviation for the intention-to-treat population (N = 145). * categories differ significantly from each other at P < 0.05.

Discussion

The results of our study provide direction in the future of obesity treatment. Our findings are consistent with prior research that show the importance of self-monitoring (in both weight and food intake) and frequent touchpoints with providers on the app for weight loss [21-24]. However, the most significant part of our study is the sustained weight loss we observed at 18-months and the behaviors associated with this long-term weight loss.

Beyond 6-months, we see certain behaviors that were crucial to initial weight loss start to drop off. While educational content and documenting food intake through photos made a significant difference between high achievers and low achievers at 6-months, these behaviors do not make an observable difference between high and low achievers at 18-months. It can be extrapolated that after 6-months, app users have reached a threshold of educational information and are capable of making good nutritional decisions on their own. They are not dependent on the educational modules nor on dietitian feedback on their food photos. These functions on the app become automated in the user's regular lifestyle and decision making, leading to a lower frequency of these behaviors in the app, while maintaining a similar level of weight loss.

As these behaviors that were essential to driving weight loss at the start of the program become statistically insignificant at the 6-month point, new behaviors emerge as better predictors of long-term weight loss [25, 26]. The difference in behaviors

between high achievers and low achievers during months 12 through months 18 provide a hint towards what is missing in the current treatment of obesity. We found that successful long-term weight loss was characterized by regular weigh-ins as well as initiating more text-based conversations. These two behaviors can be linked to long-term success because of their direct correlation with combating attrition and lack of motivation. Regular weight monitoring provides users with realistic data on their progress and helps patients avoid dramatic weight recidivism [27]. The high volume of in-app conversations emphasizes the important role that a provider-patient relationship plays in promoting behavioral changes [28-30]. Text-based conversations are a convenient way for patients to maintain touchpoints with their providers to receive personalized feedback and encouragement if they start to experience weight regain or loss of motivation.

It is also worth noting that our study utilized anti-obesity medications when deemed appropriate based on a decision made together by a medical provider and patient. While medication use did not differ significantly between high and low achievers in the 18-month cohort, it is included in our discussion to highlight the importance of multimodal treatments in a disease as complex as obesity [31, 32]. This study has the benefit of reflecting real-life situations and true clinical experiences, a perspective that randomized controlled studies do not.

Our study is unique from comparable studies on mobile engagement for obesity intervention in that it measures four different types of mobile engagement interventions: self-monitoring, feedback, support, and educational content. These distinctions help guide the future of mobile health apps by providing intervention-specific data. It also corroborates with other findings that meaningful and lasting weight loss requires a comprehensive set of interventions instead of a singular approach. Our findings suggest that mobile health applications may be better suited to achieve sustained weight loss by designing features that facilitate provider-patient interactions on top of self-tracking technology.

This study is also one of few that examines what happens to mobile engagement, and associated weight loss, over a duration of 18-months. While some studies have found short-term weight loss success through mobile app engagement, few have been able to show sustained weight loss beyond 6-months.

One limitation of our study design is the retrospective nature which does not allow us to control certain variables. In particular, the coaching in the mobile App was provided by obesity trained registered dietitians, physician assistants, and doctors. Hence, this may not be comparable to mHealth interventions where coaches are not specialty trained. Clinical expertise may be confounding the impact attributed solely through mobile coaching.

This study has demonstrated how delivering a comprehensive behavioral, pharmacologic, and nutritional weight loss program via a mobile health can help improve sustained weight loss. This is achieved by designing features within a mobile application that facilitate and elevate provider-patient relationships in order to limit and prevent weight recidivism. This study has demonstrated that technology enabled multicomponent approaches can play a significant role in the obesity treatment.

References

1. Ward ZJ, Bleich SN, Cradock AL, Barrett JL, Giles CM, et al. 2019. Projected U.S. state-level prevalence of adult obesity and severe obesity. *N Engl J Med* 381(25): 2440-2450. https://doi.org/10.1056/nejmsa1909301
2. Preston SH, Vierboom YC, Stokes A. 2018. The role of obesity in exceptionally slow US mortality improvement. *Proc Natl Acad Sci U S A* 115(5): 957-961. https://doi.org/10.1073/pnas.1716802115
3. Dalle Grave R, Claugi S, Molinari E, Petroni ML, Bondi M, et al. 2005. Weight loss expectations in obese patients and treatment attrition: an observational multicenter study. *Obes Res* 13(11): 1961-1969. https://doi.org/10.1038/oby.2005.241
4. Moroshko I, Brennan L, O'Brien P. 2011. Predictors of dropout in weight loss interventions: asystematic review of the literature. *Obes Rev* 12(11): 912-934. https://doi.org/10.1111/j.1467-789x.2011.00915.x
5. Pirotta S, Joham A, Hochberg L, Moran L, Lim S, et al. 2019. Strategies to reduce attrition in weight loss interventions: asystematic review and meta-analysis. *Obes Rev* 20(10): 1400-1412. https://doi.org/10.1111/obr.12914
6. Allen JK, Stephens J, Dennison Himmelfarb CR, Stewart KJ, Hauck S. 2013. Randomized controlled pilot study testing use of smartphone technology for obesity treatment. *J Obes* 2013: 151597. https://doi.org/10.1155/2013/151597
7. Tate DF, Wing RR, Winett RA. 2001. Using internet technology to deliver a behavioral weight loss program. *JAMA* 285(9): 1172-1177. https://doi.org/10.1001/jama.285.9.1172
8. Wharton CM, Johnston CS, Cunningham BK, Sterner D. 2014. Dietary self-monitoring, but not dietary quality, improves with use of smartphone app technology in an 8-week weight loss trial. *J Nutr Educ Behav* 46(5): 440-444. https://doi.org/10.1016/j.jneb.2014.04.291
9. Patel ML, Hopkins CM, Brooks TL, Bennett GG. 2019. Comparing self-monitoring strategies for weight loss in a smartphone app: randomized controlled trial. *JMIR Mhealth Uhealth* 7(2): e12209. https://doi.org/10.2196/12209
10. Nikolaou CK, Lean ME. 2017. Mobile applications for obesity and weight management: current market characteristics. *Int J Obes (Lond)* 41(1): 200-202. https://doi.org/10.1038/ijo.2016.186
11. Park SH, Hwang J, Choi YK. 2019. Effect of mobile health on obese adults: asystematic review and meta-analysis. *Healthc Inform Res* 25(1): 12-26. https://doi.org/10.4258/hir.2019.25.1.12
12. Bacigalupo R, Cudd P, Littlewood C, Bissell P, Hawley MS, et al. 2013. Interventions employing mobile technology for overweight and obesity: an early systematic review of randomized controlled trials. *Obes Rev* 14(4): 279-291. https://doi.org/10.1111/obr.12006
13. Laing BY, Mangione CM, Tseng CH, Leng M, Vaisberg E, et al. 2014. Effectiveness of a smartphone application for weight loss compared with usual care in overweight primary care patients: arandomized, controlled trial. *Ann Intern Med* 161(10 Suppl): S5-S12. https://doi.org/10.7326/m13-3005
14. Beleigoli AM, Andrade AQ, Cançado AG, Paulo MN, Diniz MFH. 2019. Web-based digital health interventions for weight loss and lifestyle habit changes in overweight and obese adults: systematic review and meta-analysis. *J Med Internet Res* 21(1): e298. https://doi.org/10.2196/jmir.9609
15. Khokhar B, Jones J, Ronksley PE, Armstrong MJ, Caird J, et al. 2014. Effectiveness of mobile electronic devices in weight loss among overweight and obese populations: asystematic review and meta-analysis. *BMC Obes* 1: 22. https://doi.org/10.1186/s40608-014-0022-4
16. Chin SO, Keum C, Woo J, Park J, Choi HJ, et al. 2016. Successful

weight reduction and maintenance by using asmartphone application in those with overweight and obesity. *Sci Rep* 6: 34563. https://doi.org/10.1038/srep34563

17. Ferrara G, Kim J, Lin S, Hua J, Seto E. 2019. A focused review of smartphone diet-tracking apps: usability, functionality, coherence with behavior change theory, and comparative validity of nutrient intake and energy estimates. *JMIR Mhealth Uhealth* 7(5): e9232. https://doi.org/10.2196/mhealth.9232
18. Kankanhalli A, Shin J, Oh H. 2019. Mobile-based interventions for dietary behavior change and health outcomes: scoping review. *JMIR Mhealth Uhealth* 7(1): e11312. https://doi.org/10.2196/11312
19. Bardus M, van Beurden SB, Smith JR, Abraham C. 2016. A review and content analysis of engagement, functionality, aesthetics, information quality, and change techniques in the most popular commercial apps for weight management. *Int J Behav Nutr Phys Act* 13: 35. https://doi.org/10.1186/s12966-016-0359-9
20. Painter S, Ditsch G, Ahmed R, Hanson NB, Kachin K, et al. 2016. Retrofit weight-loss outcomes at 6, 12, and 24 months and characteristics of 12-month high performers: aretrospective analysis. *JMIR Mhealth Uhealth* 4(3): e101. https://doi.org/10.2196/mhealth.5873
21. Harvey J, Krukowski R, Priest J, West D. 2019. Log often, lose more: electronic dietary self-monitoring for weight loss. *Obesity (Silver Spring)* 27(3): 380-384. https://doi.org/10.1002/oby.22382
22. Peterson ND, Middleton KR, Nackers LM, Medina KE, Milsom VA, et al. 2014. Dietary self-monitoring and long-term success with weight management. *Obesity (Silver Spring)* 22(9): 1962-1967. https://doi.org/10.1002/oby.20807
23. Rose SA, Poynter PS, Anderson JW, Noar SM, Conigliaro J. 2013. Physician weight loss advice and patient weight loss behavior change: aliterature review and meta-analysis of survey data. *Int J Obes (Lond)* 37(1): 118-128. https://doi.org/10.1038/ijo.2012.24
24. McVay M, Steinberg D, Askew S, Bennett GG. 2019. Provider counseling and weight loss outcomes in a primary care-based digital obesity treatment. *J Gen Intern Med* 34(6): 992-998. https://doi.org/10.1007/s11606-019-04944-5
25. Sciamanna CN, Kiernan M, Rolls BJ, Boan J, Stuckey H et al. 2011. Practices associated with weight loss versus weight-loss maintenance: results of anational survey. *Am J Prev Med* 41(2): 159-166. https://doi.org/10.1016/j.amepre.2011.04.009
26. Nakade M, Aiba N, Morita A, Miyachi M, Sasaki S, et al. 2012. What behaviors are important for successful weight maintenance? *J Obes* 2012: 202037. https://doi.org/10.1155/2012/202037
27. Byrne S, Cooper Z, Fairburn C. 2003. Weight maintenance and relapse in obesity: aqualitative study. *Int J Obes Relat Metab Disord* 27(8): 955-962. https://doi.org/10.1038/sj.ijo.0802305
28. Mayor S. 2016. GP referral to weight loss programme is quick and effective, finds study. *BMJ* 355: i5727. https://doi.org/10.1136/bmj.i5727
29. Jordan K. 2019. A provider-patient relationship: the critical first step of smoking cessation. *J Gen Intern Med* 35(1): 10-11. https://doi.org/10.1007/s11606-019-05276-0
30. Kahan SI. 2018. Practical strategies for engaging individuals with obesity in primary care. *Mayo Clin Proc* 93(3): 351-359. https://doi.org/10.1016/j.mayocp.2018.01.006
31. Taylor J. 2020. Looking beyond lifestyle: acomprehensive approach to the treatment of obesity in the primary care setting. *JNP-J Nurse Pract* 16(1): 74-78. https://doi.org/10.1016/j.nurpra.2019.09.021
32. Lavie CJ, Laddu D, Arena R, Ortega FB, Alpert MA, et al. 2018. Healthy weight and obesity prevention: JACC health promotion series. *J Am Coll Cardiol* 72(13): 1506-1531. https://doi.org/10.1016/j.jacc.2018.08.1037

Bibliography

[1] Madhumita Basu and Vidu Garg. Maternal hyperglycemia and fetal cardiac development: Clinical impact and underlying mechanisms. *Birth Defects Research*, 110, 2018.

[2] R. Safavi, A. Lih, S. Kirkpatrick, S. Haller, and M. R. Bailony. Impact of anti-obesity medication initiation and duration on weight loss in a comprehensive weight loss programme. *Obesity Science and Practice*, 5, 2019.

[3] Hung Nguyen, Duc Tran, Bang Tran, Bahadir Pehlivan, and Tin Nguyen. A comprehensive survey of regulatory network inference methods using single cell rna sequencing data. *Briefings in Bioinformatics*, 22, 2021.

[4] Yoonjee Kang, Denis Thieffry, and Laura Cantini. Evaluating the reproducibility of single-cell gene regulatory network inference algorithms. *Frontiers in Genetics*, 12, 2021.

[5] Frank Emmert-Streib, Matthias Dehmer, and Benjamin Haibe-Kains. Gene regulatory networks and their applications: Understanding biological and medical

problems in terms of networks. *Frontiers in Cell and Developmental Biology*, 2, 2014.

[6] M. Oti, B. Snel, M. A. Huynen, and Han G. Brunner. Predicting disease genes using protein-protein interactions. *Journal of Medical Genetics*, 43, 2006.

[7] Trey Ideker and Roded Sharan. Protein networks in disease. *Genome Research*, 18, 2008.

[8] Katia Basso, Adam A. Margolin, Gustavo Stolovitzky, Ulf Klein, Riccardo Dalla-Favera, and Andrea Califano. Reverse engineering of regulatory networks in human b cells. *Nature Genetics*, 37, 2005.

[9] Antoine Emmanuel Saliba, Alexander J. Westermann, Stanislaw A. Gorski, and Jörg Vogel. Single-cell rna-seq: Advances and future challenges. *Nucleic Acids Research*, 42, 2014.

[10] C. Wyatt Shields Iv, Catherine D. Reyes, and Gabriel P. López. Microfluidic cell sorting: A review of the advances in the separation of cells from debulking to rare cell isolation. *Lab on a Chip*, 15, 2015.

[11] Aditya Pratapa, Amogh P. Jalihal, Jeffrey N. Law, Aditya Bharadwaj, and T. M. Murali. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature Methods*, 17, 2020.

[12] Haochang Hu, Shaoyi Lin, Shuangshuang Wang, and Xiaomin Chen. The role

of transcription factor 21 in epicardial cell differentiation and the development of coronary heart disease. *Frontiers in Cell and Developmental Biology*, 8, 2020.

[13] Kyle Akers and T. M. Murali. Gene regulatory network inference in single-cell biology. *Current Opinion in Systems Biology*, 26, 2021.

[14] Nan Papili Gao, S. M.Minhaz Ud-Dean, Olivier Gandrillon, and Rudiyanto Gunawan. Sincerities: Inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. *Bioinformatics*, 34, 2018.

[15] Vân Anh Huynh-Thu, Alexandre Irrthum, Louis Wehenkel, and Pierre Geurts. Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE*, 5, 2010.

[16] Shuonan Chen and Jessica C. Mar. Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC Bioinformatics*, 19, 2018.

[17] Sara Aibar, Carmen Bravo González-Blas, Thomas Moerman, Vân Anh Huynh-Thu, Hana Imrichova, Gert Hulselmans, Florian Rambow, Jean Christophe Marine, Pierre Geurts, Jan Aerts, Joost Van Den Oord, Zeynep Kalender Atak, Jasper Wouters, and Stein Aerts. Scenic: Single-cell regulatory network inference and clustering. *Nature Methods*, 14, 2017.

[18] François Spitz. Gene regulation at a distance: From remote enhancers to 3d regulatory ensembles. *Seminars in Cell and Developmental Biology*, 57, 2016.

[19] Evgeny Z. Kvon, Tomas Kazmar, Gerald Stampfel, J. Omar Yáñez-Cuna, Michaela Pagani, Katharina Schernhuber, Barry J. Dickson, and Alexander Stark. Genome-scale functional characterization of drosophila developmental enhancers in vivo. *Nature*, 512, 2014.

[20] Chun Ping Yu, Jinn Jy Lin, and Wen Hsiung Li. Positional distribution of transcription factor binding sites in arabidopsis thaliana. *Scientific Reports*, 6, 2016.

[21] Rémi Xavier Coux, Nick D.L. Owens, and Pablo Navarro. Chromatin accessibility and transcription factor binding through the perspective of mitosis. *Transcription*, 11, 2020.

[22] Bing Li, Michael Carey, and Jerry L. Workman. The role of chromatin during transcription. *Cell*, 128, 2007.

[23] Jason D. Buenrostro, Paul G. Giresi, Lisa C. Zaba, Howard Y. Chang, and William J. Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, dna-binding proteins and nucleosome position. *Nature Methods*, 10, 2013.

[24] Jason D. Buenrostro, Beijing Wu, Howard Y. Chang, and William J. Greenleaf. Atac-seq: A method for assaying chromatin accessibility genome-wide. *Current Protocols in Molecular Biology*, 2015, 2015.

[25] Badia i Mompel P., Vélez Santiago J., Braunger J., Geiss C., Dimitrov D., Müller-Dott S., Taus P., Holland C.H. Dugourd A., Ramirez Flores R.O., and Saez-

Rodriguez J. decoupler: ensemble of computational methods to infer biological activities from omics data, bioinformatics advances. *Bioinformatics Advances*, 2, 2022.

[26] Isabelle S. Peter. Methods for the experimental and computational analysis of gene regulatory networks in sea urchins, 2019.

[27] Joel Smith, Christina Theodoris, and Eric H. Davidson. A gene regulatory network subcircuit drives a dynamic pattern of gene expression. *Science*, 318, 2007.

[28] Eric H. Davidson, Jonathan P. Rast, Paola Oliveri, Andrew Ransick, Cristina Calestani, Chiou Hwa Yuh, Takuya Minokawa, Gabriele Amore, Veronica Hinman, César Arenas-Mena, Ochan Otim, C. Titus Brown, Carolina B. Livi, Pei Yun Lee, Roger Revilla, Alistair G. Rust, Zheng Jun Pan, Maria J. Schilstra, Peter J.C. Clarke, Maria I. Arnone, Lee Rowen, R. Andrew Cameron, David R. McClay, Leroy Hood, and Hamid Bolouri. A genomic regulatory network for development. *Science*, 295, 2002.

[29] Isabelle S. Peter and Eric H. Davidson. *Genomic Control Process: Development and Evolution*. 2015.

[30] Lesley T. MacNeil and Albertha J.M. Walhout. Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression. *Genome Research*, 21, 2011.

[31] Lingyao Zeng, Husain A. Talukdar, Simon Koplev, Chiara Giannarelli, Torbjörn

Ivert, Li Ming Gan, Arno Ruusalepp, Eric E. Schadt, Jason C. Kovacic, Aldons J. Lusis, Tom Michoel, Heribert Schunkert, and Johan L.M. Björkegren. Contribution of gene regulatory networks to heritability of coronary artery disease. *Journal of the American College of Cardiology*, 73, 2019.

[32] Irfan S. Kathiriya, Kavitha S. Rao, Giovanni Iacono, W. Patrick Devine, Andrew P. Blair, Swetansu K. Hota, Michael H. Lai, Bayardo I. Garay, Reuben Thomas, Henry Z. Gong, Lauren K. Wasson, Piyush Goyal, Tatyana Sukonnik, Kevin M. Hu, Gunes A. Akgun, Laure D. Bernard, Brynn N. Akerberg, Fei Gu, Kai Li, Matthew L. Speir, Maximilian Haeussler, William T. Pu, Joshua M. Stuart, Christine E. Seidman, J. G. Seidman, Holger Heyn, and Benoit G. Bruneau. Modeling human tbx5 haploinsufficiency predicts regulatory networks for congenital heart disease. *Developmental Cell*, 56, 2021.

[33] Altuna Akalin. *Computational Genomics with R.* 2020.

[34] Len A. Pennacchio, Wendy Bickmore, Ann Dean, Marcelo A. Nobrega, and Gill Bejerano. Enhancers: Five essential questions. *Nature Reviews Genetics*, 14, 2013.

[35] Vân Anh Huynh-Thu, Alexandre Irrthum, Louis Wehenkel, and Pierre Geurts. Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE*, 5, 2010.

[36] Federico Zambelli, Graziano Pesole, and Giulio Pavesi. Motif discovery and tran-

scription factor binding sites before and after the next-generation sequencing era. *Briefings in Bioinformatics*, 14, 2013.

[37] Gonzalo Millán-Zambrano, Adam Burton, Andrew J. Bannister, and Robert Schneider. Histone post-translational modifications — cause and consequence of genome function. *Nature Reviews Genetics*, 2022.

[38] Heather Stegenga, Alexander Haines, Katie Jones, and John Wilding. Identification, assessment, and management of overweight and obesity: summary of updated nice guidance. *BMJ*, 349, 2014.

[39] Fatma A. Hashim, Mai S. Mabrouk, and Walid Al-Atabany. Review of different sequence motif finding algorithms. *Avicenna Journal of Medical Biotechnology*, 11, 2019.

[40] Timothy L. Bailey and Charles Elkan. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning*, 21, 1995.

[41] Benjamin L. Kidder, Gangqing Hu, and Keji Zhao. Chip-seq: Technical considerations for obtaining high-quality data. *Nature Immunology*, 12, 2011.

[42] E. Christopher Partridge, Surya B. Chhetri, Jeremy W. Prokop, Ryne C. Ramaker, Camden S. Jansen, Say Tar Goh, Mark Mackiewicz, Kimberly M. Newberry, Laurel A. Brandsmeier, Sarah K. Meadows, C. Luke Messer, Andrew A. Hardigan, Candice J. Coppola, Emma C. Dean, Shan Jiang, Daniel Savic, Ali Mortazavi, Barbara J. Wold, Richard M. Myers, and Eric M. Mendenhall. Occu-

pancy maps of 208 chromatin-associated proteins in one human cell type. *Nature*, 583, 2020.

[43] Cricket A. Sloan, Esther T. Chan, Jean M. Davidson, Venkat S. Malladi, J. Seth Strattan, Benjamin C. Hitz, Idan Gabdank, Aditi K. Narayanan, Marcus Ho, Brian T. Lee, Laurence D. Rowe, Timothy R. Dreszer, Greg Roe, Nikhil R. Podduturi, Forrest Tanaka, Eurie L. Hong, and J. Michael Cherry. Encode data at the encode portal. *Nucleic Acids Research*, 44, 2016.

[44] John S. Satterlee, Lisa H. Chadwick, Frederick L. Tyson, Kim McAllister, Jill Beaver, Linda Birnbaum, Nora D. Volkow, Elizabeth L. Wilder, James M. Anderson, and Ananda L. Roy. The nih common fund/roadmap epigenomics program: Successes of a comprehensive consortium. *Science Advances*, 5, 2019.

[45] Byungjin Hwang, Ji Hyun Lee, and Duhee Bang. Single-cell rna sequencing technologies and bioinformatics pipelines. *Experimental and Molecular Medicine*, 50, 2018.

[46] Zhicheng Ji, Weiqiang Zhou, Wenpin Hou, and Hongkai Ji. Single-cell atac-seq signal extraction and enhancement with scate. *Genome Biology*, 21, 2020.

[47] Aleksandra A. Kolodziejczyk, Jong Kyoung Kim, Valentine Svensson, John C. Marioni, and Sarah A. Teichmann. The technology and biology of single-cell rna sequencing. *Molecular Cell*, 58, 2015.

[48] Sebastian Pott and Jason D. Lieb. Single-cell atac-seq: Strength in numbers. *Genome Biology*, 16, 2015.

[49] Author Manuscript. Transposition of native chromatin for multimodal regulatory analysis and personal epigenomics. *Biochemistry Journal*, 416, 2008.

[50] Carmen Bravo González-Blas, Xiao-Jiang Quan, Ramon Duran-Romaña, Ibrahim Ihsan Taskiran, Duygu Koldere, Kristofer Davie, Valerie Christiaens, Samira Makhzami, Gert Hulselmans, Maxime Waegeneer, David Mauduit, Suresh Poovathingal, Sara Aibar, and Stein Aerts. Identification of genomic enhancers through spatial integration of single-cell transcriptomics and epigenomics. *Molecular Systems Biology*, 16, 2020.

[51] Haofei Wang, Yuchen Yang, Yunzhe Qian, Jiandong Liu, and Li Qian. Delineating chromatin accessibility re-patterning at single cell level during early stage of direct cardiac reprogramming. *Journal of Molecular and Cellular Cardiology*, 162, 2022.

[52]

[53] Daniel Marbach, James C. Costello, Robert Küffner, Nicole M. Vega, Robert J. Prill, Diogo M. Camacho, Kyle R. Allison, Manolis Kellis, James J. Collins, Andrej Aderhold, Gustavo Stolovitzky, Richard Bonneau, Yukun Chen, Francesca Cordero, Martin Crane, Frank Dondelinger, Mathias Drton, Roberto Esposito, Rina Foygel, Alberto De La Fuente, Jan Gertheiss, Pierre Geurts, Alex Greenfield, Marco Grzegorczyk, Anne Claire Haury, Benjamin Holmes, Torsten Hothorn,

Dirk Husmeier, Vân Anh Huynh-Thu, Alexandre Irrthum, Guy Karlebach, Sophie Lèbre, Vincenzo De Leo, Aviv Madar, Subramani Mani, Fantine Mordelet, Harry Ostrer, Zhengyu Ouyang, Ravi Pandya, Tobias Petri, Andrea Pinna, Christopher S. Poultney, Serena Rezny, Heather J. Ruskin, Yvan Saeys, Ron Shamir, Alina Sîrbu, Mingzhou Song, Nicola Soranzo, Alexander Statnikov, Nicci Vega, Paola Vera-Licona, Jean Philippe Vert, Alessia Visconti, Haizhou Wang, Louis Wehenkel, Lukas Windhager, Yang Zhang, and Ralf Zimmer. Wisdom of crowds for robust gene network inference. *Nature Methods*, 9, 2012.

[54] Feng Yan, David R. Powell, David J. Curtis, and Nicholas C. Wong. From reads to insight: A hitchhiker's guide to atac-seq data analysis, 2020.

[55] Hannah A. Pliner, Jonathan S. Packer, José L. McFaline-Figueroa, Darren A. Cusanovich, Riza M. Daza, Delasa Aghamirzaie, Sanjay Srivatsan, Xiaojie Qiu, Dana Jackson, Anna Minkina, Andrew C. Adey, Frank J. Steemers, Jay Shendure, and Cole Trapnell. Cicero predicts cis-regulatory dna interactions from single-cell chromatin accessibility data. *Molecular Cell*, 71, 2018.

[56] Amanda M. Ackermann, Zhiping Wang, Jonathan Schug, Ali Naji, and Klaus H. Kaestner. Integration of atac-seq and rna-seq identifies human alpha cell and beta cell signature genes. *Molecular Metabolism*, 5, 2016.

[57] Jing Qin, Yaohua Hu, Feng Xu, Hari Krishna Yalamanchili, and Junwen Wang. Inferring gene regulatory networks by integrating chip-seq/chip and transcriptome data via lasso-type regularization methods. *Methods*, 67, 2014.

[58] Junyue Cao, Darren A. Cusanovich, Vijay Ramani, Delasa Aghamirzaie, Hannah A. Pliner, Andrew J. Hill, Riza M. Daza, Jose L. McFaline-Figueroa, Jonathan S. Packer, Lena Christiansen, Frank J. Steemers, Andrew C. Adey, Cole Trapnell, and Jay Shendure. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*, 361, 2018.

[59] Comprehensive integration of single-cell data. *Cell*, 177, 2019.

[60] Archr is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nature Genetics*, 53, 2021.

[61] Nicholas Pervolarakis, Quy H. Nguyen, Justice Williams, Yanwen Gong, Guadalupe Gutierrez, Peng Sun, Darisha Jhutty, Grace X.Y. Zheng, Corey M. Nemec, Xing Dai, Kazuhide Watanabe, and Kai Kessenbrock. Integrated single-cell transcriptomics and chromatin accessibility analysis reveals regulators of mammary epithelial cell identity. *Cell Reports*, 33, 2020.

[62] Jeffrey M. Granja, Sandy Klemm, Lisa M. McGinnis, Arwa S. Kathiria, Anja Mezger, M. Ryan Corces, Benjamin Parks, Eric Gars, Michaela Liedtke, Grace X.Y. Zheng, Howard Y. Chang, Ravindra Majeti, and William J. Greenleaf. Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nature Biotechnology*, 37, 2019.

[63] Alexandro E. Trevino, Fabian Müller, Jimena Andersen, Laksshman Sundaram, Arwa Kathiria, Anna Shcherbina, Kyle Farh, Howard Y. Chang, Anca M. Paşca,

Anshul Kundaje, Sergiu P. Pașca, and William J. Greenleaf. Chromatin and gene-regulatory dynamics of the developing human cerebral cortex at single-cell resolution. *Cell*, 184, 2021.

[64] Mattia Forcato, Oriana Romano, and Silvio Bicciato. Computational methods for the integrative analysis of single-cell data. *Briefings in Bioinformatics*, 22, 2021.

[65] Federico Abascal, Reyes Acosta, Nicholas J. Addleman, Jessika Adrian, Veena Afzal, Bronwen Aken, Jennifer A. Akiyama, Omar Al Jammal, Henry Amrhein, Stacie M. Anderson, Gregory R. Andrews, Igor Antoshechkin, Kristin G. Ardlie, Joel Armstrong, Matthew Astley, Budhaditya Banerjee, Amira A. Barkal, If H.A. Barnes, Iros Barozzi, Daniel Barrell, Gemma Barson, Daniel Bates, Ulugbek K. Baymuradov, Cassandra Bazile, Michael A. Beer, Samantha Beik, M. A. Bender, Ruth Bennett, Louis Philip Benoit Bouvrette, Bradley E. Bernstein, Andrew Berry, Anand Bhaskar, Alexandra Bignell, Steven M. Blue, David M. Bodine, Carles Boix, Nathan Boley, Tyler Borrman, Beatrice Borsari, Alan P. Boyle, Laurel A. Brandsmeier, Alessandra Breschi, Emery H. Bresnick, Jason A. Brooks, Michael Buckley, Christopher B. Burge, Rachel Byron, Eileen Cahill, Lingling Cai, Lulu Cao, Mark Carty, Rosa G. Castanon, Andres Castillo, Hassan Chaib, Esther T. Chan, Daniel R. Chee, Sora Chee, Hao Chen, Huaming Chen, Jia Yu Chen, Songjie Chen, J. Michael Cherry, Surya B. Chhetri, Jyoti S. Choudhary, Jacqueline Chrast, Dongjun Chung, Declan Clarke, Neal A.L. Cody, Candice J. Coppola, Julie Coursen, Anthony M. D'Ippolito, Stephen Dalton, Cassidy Danyko,

Claire Davidson, Jose Davila-Velderrain, Carrie A. Davis, Job Dekker, Alden Deran, Gilberto DeSalvo, Gloria Despacio-Reyes, Colin N. Dewey, Diane E. Dickel, Morgan Diegel, Mark Diekhans, Vishnu Dileep, Bo Ding, Sarah Djebali, Alexander Dobin, Daniel Dominguez, Sarah Donaldson, Jorg Drenkow, Timothy R. Dreszer, Yotam Drier, Michael O. Duff, Douglass Dunn, Catharine Eastman, Joseph R. Ecker, Matthew D. Edwards, Nicole El-Ali, Shaimae I. Elhajjajy, Keri Elkins, Andrew Emili, Charles B. Epstein, Rachel C. Evans, Iakes Ezkurdia, Kaili Fan, Peggy J. Farnham, Nina Farrell, Elise A. Feingold, Anne Maud Ferreira, Katherine Fisher-Aylor, Stephen Fitzgerald, Paul Flicek, Chuan Sheng Foo, Kevin Fortier, Adam Frankish, Peter Freese, Shaliu Fu, Xiang Dong Fu, Yu Fu, Yoko Fukuda-Yuzawa, Mariateresa Fulciniti, Alister P.W. Funnell, Idan Gabdank, Timur Galeev, Mingshi Gao, Carlos Garcia Giron, Tyler H. Garvin, Chelsea Anne Gelboin-Burkhart, Grigorios Georgolopoulos, Mark B. Gerstein, Belinda M. Giardine, David K. Gifford, David M. Gilbert, Daniel A. Gilchrist, Shawn Gillespie, Thomas R. Gingeras, Peng Gong, Alvaro Gonzalez, Jose M. Gonzalez, Peter Good, Alon Goren, David U. Gorkin, Brenton R. Graveley, Michael Gray, Jack F. Greenblatt, Ed Griffiths, Mark T. Groudine, Fabian Grubert, Mengting Gu, Roderic Guigó, Hongbo Guo, Yu Guo, Yuchun Guo, Gamze Gursoy, Maria Gutierrez-Arcelus, Jessica Halow, Ross C. Hardison, Matthew Hardy, Manoj Hariharan, Arif Harmanci, Anne Harrington, Jennifer L. Harrow, Tatsunori B. Hashimoto, Richard D. Hasz, Meital Hatan, Eric Haugen, James E. Hayes, Peng He, Yupeng He, Nastaran Heidari, David Hendrickson, Elisabeth F. Heuston, Jason A.

Hilton, Benjamin C. Hitz, Abigail Hochman, Cory Holgren, Lei Hou, Shuyu Hou, Yun Hua E. Hsiao, Shanna Hsu, Hui Huang, Tim J. Hubbard, Jack Huey, Timothy R. Hughes, Toby Hunt, Sean Ibarrientos, Robbyn Issner, Mineo Iwata, Osagie Izuogu, Tommi Jaakkola, Nader Jameel, Camden Jansen, Lixia Jiang, Peng Jiang, Audra Johnson, Rory Johnson, Irwin Jungreis, Madhura Kadaba, Maya Kasowski, Mary Kasparian, Momoe Kato, Rajinder Kaul, Trupti Kawli, Michael Kay, Judith C. Keen, Sunduz Keles, Cheryl A. Keller, David Kelley, Manolis Kellis, Pouya Kheradpour, Daniel Sunwook Kim, Anthony Kirilusha, Robert J. Klein, Birgit Knoechel, Samantha Kuan, Michael J. Kulik, Sushant Kumar, Anshul Kundaje, Tanya Kutyavin, Julien Lagarde, Bryan R. Lajoie, Nicole J. Lambert, John Lazar, Ah Young Lee, Donghoon Lee, Elizabeth Lee, Jin Wook Lee, Kristen Lee, Christina S. Leslie, Shawn Levy, Bin Li, Hairi Li, Nan Li, Xiangrui Li, Yang I. Li, Ying Li, Yining Li, Yue Li, Jin Lian, Maxwell W. Libbrecht, Shin Lin, Ying Lin, Dianbo Liu, Jason Liu, Peng Liu, Tingting Liu, X. Shirley Liu, Yan Liu, Yaping Liu, Maria Long, Shaoke Lou, Jane Loveland, Aiping Lu, Yuheng Lu, Eric Lécuyer, Lijia Ma, Mark Mackiewicz, Brandon J. Mannion, Michael Mannstadt, Deepa Manthravadi, Georgi K. Marinov, Fergal J. Martin, Eugenio Mattei, Kenneth McCue, Megan McEown, Graham McVicker, Sarah K. Meadows, Alex Meissner, Eric M. Mendenhall, Christopher L. Messer, Wouter Meuleman, Clifford Meyer, Steve Miller, Matthew G. Milton, Tejaswini Mishra, Dianna E. Moore, Helen M. Moore, Jill E. Moore, Samuel H. Moore, Jennifer Moran, Ali Mortazavi, Jonathan M. Mudge, Nikhil Munshi, Rabi Murad, Richard M. Myers,

Vivek Nandakumar, Preetha Nandi, Anil M. Narasimha, Aditi K. Narayanan, Hannah Naughton, Fabio C.P. Navarro, Patrick Navas, Jurijs Nazarovs, Jemma Nelson, Shane Neph, Fidencio Jun Neri, Joseph R. Nery, Amy R. Nesmith, J. Scott Newberry, Kimberly M. Newberry, Vu Ngo, Rosy Nguyen, Thai B. Nguyen, Tung Nguyen, Andrew Nishida, William S. Noble, Catherine S. Novak, Eva Maria Novoa, Briana Nuñez, Charles W. O'Donnell, Sara Olson, Kathrina C. Onate, Ericka Otterman, Hakan Ozadam, Michael Pagan, Tsultrim Palden, Xinghua Pan, Yongjin Park, E. Christopher Partridge, Benedict Paten, Florencia Pauli-Behn, Michael J. Pazin, Baikang Pei, Len A. Pennacchio, Alexander R. Perez, Emily H. Perry, Dmitri D. Pervouchine, Nishigandha N. Phalke, Quan Pham, Doug H. Phanstiel, Ingrid Plajzer-Frick, Gabriel A. Pratt, Henry E. Pratt, Sebastian Preissl, Jonathan K. Pritchard, Yuri Pritykin, Michael J. Purcaro, Qian Qin, Giovanni Quinones-Valdez, Ines Rabano, Ernest Radovani, Anil Raj, Nisha Rajagopal, Oren Ram, Lucia Ramirez, Ricardo N. Ramirez, Dylan Rausch, Soumya Raychaudhuri, Joseph Raymond, Rozita Razavi, Timothy E. Reddy, Thomas M. Reimonn, Bing Ren, Alexandre Reymond, Alex Reynolds, Suhn K. Rhie, John Rinn, Miguel Rivera, Juan Carlos Rivera-Mulia, Brian Roberts, Jose Manuel Rodriguez, Joel Rozowsky, Russell Ryan, Eric Rynes, Denis N. Salins, Richard Sandstrom, Takayo Sasaki, Shashank Sathe, Daniel Savic, Alexandra Scavelli, Jonathan Scheiman, Christoph Schlaffner, Jeffery A. Schloss, Frank W. Schmitges, Lei Hoon See, Anurag Sethi, Manu Setty, Anthony Shafer, Shuo Shan, Eilon Sharon, Quan Shen, Yin Shen, Richard I. Sherwood, Minyi Shi, Sunyoung Shin,

Noam Shoresh, Kyle Siebenthall, Cristina Sisu, Teri Slifer, Cricket A. Sloan, Anna Smith, Valentina Snetkova, Michael P. Snyder, Damek V. Spacek, Sharanya Srinivasan, Rohith Srivas, George Stamatoyannopoulos, John A. Stamatoyannopoulos, Rebecca Stanton, Dave Steffan, Sandra Stehling-Sun, J. Seth Strattan, Amanda Su, Balaji Sundararaman, Marie Marthe Suner, Tahin Syed, Matt Szynkarek, Forrest Y. Tanaka, Danielle Tenen, Mingxiang Teng, Jeffrey A. Thomas, Dave Toffey, Michael L. Tress, Diane E. Trout, Gosia Trynka, Junko Tsuji, Sean A. Upchurch, Oana Ursu, Barbara Uszczynska-Ratajczak, Mia C. Uziel, Alfonso Valencia, Benjamin Van Biber, Arjan G. van der Velde, Eric L. Van Nostrand, Yekaterina Vaydylevich, Jesus Vazquez, Alec Victorsen, Jost Vielmetter, Jeff Vierstra, Axel Visel, Anna Vlasova, Christopher M. Vockley, Simona Volpi, Shinny Vong, Hao Wang, Mengchi Wang, Qin Wang, Ruth Wang, Tao Wang, Wei Wang, Xiaofeng Wang, Yanli Wang, Nathaniel K. Watson, Xintao Wei, Zhijie Wei, Hendrik Weisser, Sherman M. Weissman, Rene Welch, Robert E. Welikson, Zhiping Weng, Harm Jan Westra, John W. Whitaker, Collin White, Kevin P. White, Andre Wildberg, Brian A. Williams, David Wine, Heather N. Witt, Barbara Wold, Maxim Wolf, James Wright, Rui Xiao, Xinshu Xiao, Jie Xu, Jinrui Xu, Koon Kiu Yan, Yongqi Yan, Hongbo Yang, Xinqiong Yang, Yi Wen Yang, Galip Gürkan Yardımcı, Brian A. Yee, Gene W. Yeo, Taylor Young, Tianxiong Yu, Feng Yue, Chris Zaleski, Chongzhi Zang, Haoyang Zeng, Weihua Zeng, Daniel R. Zerbino, Jie Zhai, Lijun Zhan, Ye Zhan, Bo Zhang, Jialing Zhang, Jing Zhang, Kai Zhang, Lijun Zhang, Peng Zhang, Qi Zhang, Xiao Ou Zhang, Yanxiao Zhang, Zhizhuo Zhang, Yuan

Zhao, Ye Zheng, Guoqing Zhong, Xiao Qiao Zhou, Yun Zhu, Jared Zimmerman, Michael P. Snyder, Thomas R. Gingeras, Jill E. Moore, Zhiping Weng, Mark B. Gerstein, Bing Ren, Ross C. Hardison, John A. Stamatoyannopoulos, Brenton R. Graveley, Elise A. Feingold, Michael J. Pazin, Michael Pagan, Daniel A. Gilchrist, Benjamin C. Hitz, J. Michael Cherry, Bradley E. Bernstein, Eric M. Mendenhall, Daniel R. Zerbino, Adam Frankish, Paul Flicek, and Richard M. Myers. Perspectives on encode, 2020.

[66] James D. Hocker, Olivier B. Poirion, Fugui Zhu, Justin Buchanan, Kai Zhang, Joshua Chiou, Tsui Min Wang, Qingquan Zhang, Xiaomeng Hou, Yang E. Li, Yanxiao Zhang, Elie N. Farah, Allen Wang, Andrew D. McCulloch, Kyle J. Gaulton, Bing Ren, Neil C. Chi, and Sebastian Preissl. Cardiac cell type-specific gene regulatory programs and disease risk association. *Science Advances*, 7, 2021.

[67] Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36, 2018.

[68] Yong Zhang, Tao Liu, Clifford A. Meyer, Jérôme Eeckhoute, David S. Johnson, Bradley E. Bernstein, Chad Nussbaum, Richard M. Myers, Myles Brown, Wei Li, and X. Shirley Shirley. Model-based analysis of chip-seq (macs). *Genome Biology*, 9, 2008.

[69] Daniela M. Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix

decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10, 2009.

[70] Maria Tsompana and Michael J. Buck. Chromatin accessibility: A window into the genome. *Epigenetics and Chromatin*, 7, 2014.

[71] PBMCs from GRCh38 healthy donor (v3, 150x150), single cell immune profiling dataset by cell ranger 3.0.0, 10x genomics, (2018, november 19).

[72] Grace J. Liu, Markus Jaritz, Miriam Wöhner, Benedikt Agerer, Andreas Bergthaler, Stephen G. Malin, and Meinrad Busslinger. Repression of the b cell identity factor pax5 is not required for plasma cell development. *Journal of Experimental Medicine*, 217, 2020.

[73] Jihong Dai, Nicholas J. Megjugorac, Sheela B. Amrute, and Patricia Fitzgerald-Bocarsly. Regulation of ifn regulatory factor-7 and ifn- production by enveloped virus and lipopolysaccharide in human plasmacytoid dendritic cells. *The Journal of Immunology*, 173, 2004.

[74] Brian Webster, Scott W. Werneke, Biljana Zafirova, Sébastien This, Séverin Coléon, Elodie Décembre, Helena Paidassi, Isabelle Bouvier, Pierre Emmanuel Joubert, Darragh Duffy, Thierry Walzer, Matthew L. Albert, and Marlène Dreux. Plasmacytoid dendritic cells control dengue and chikungunya virus infections via irf7-regulated interferon responses. *eLife*, 7, 2018.

[75] Lu Yang, Yanbing Zhu, Dan Tian, Song Wang, Jincheng Guo, Guangyong Sun,

Hua Jin, Chunpan Zhang, Wen Shi, M. Eric Gershwin, Zhongtao Zhang, Yi Zhao, and Dong Zhang. Transcriptome landscape of double negative t cells by single-cell rna sequencing. *Journal of Autoimmunity*, 121, 2021.

[76] Zheng Pan, Christopher J. Hetherington, and Dong-Er Zhang. Ccaat/enhancer-binding protein activates the cd14 promoter and mediates transforming growth factor signaling in monocyte development*. *Journal of Biological Chemistry*, 274(33):23242–23248, 1999.

[77] C. Arnold Spek, Hella L. Aberson, Joe M. Butler, Alex F. de Vos, and Janwillem Duitman. Cebpd potentiates the macrophage inflammatory response but cebpd knock-out macrophages fail to identify cebpd-dependent pro-inflammatory transcriptional programs. *Cells*, 10, 2021.

[78] Ditsa Levanon, Varda Negreanu, Joseph Lotem, Karen Rae Bone, Ori Brenner, Dena Leshkowitz, and Yoram Groner. Transcription factor runx3 regulates interleukin-15-dependent natural killer cell activation. *Molecular and Cellular Biology*, 34, 2014.

[79] C. Benjamin Lai and Dixie L. Mager. Role of runt-related transcription factor 3 (runx3) in transcription regulation of natural cytotoxicity receptor 1 (ncr1/nkp46), an activating natural killer (nk) cell receptor. *Journal of Biological Chemistry*, 287, 2012.

[80] Xin Huang, Cunte Chen, Mengjun Zhong, Suxia Geng, Yujie Zhao, Minming Li,

Chenxin Deng, Lingji Zeng, Ping Wu, Zesheng Lu, Jianyu Weng, Xin Du, and Yangqiu Li. Lower bcl11b expression is associated with adverse clinical outcome for patients with myelodysplastic syndrome. *Biomarker Research*, 9, 2021.

[81] Cheryl Mei Jun Tan and Adam James Lewandowski. The transitional heart: From early embryonic and fetal development to neonatal life. *Fetal Diagnosis and Therapy*, 47, 2020.

[82] Weiliang Wu, Jinxian He, and Xiaobo Shao. Incidence and mortality trend of congenital heart disease at the global, regional, and national level, 1990-2017. *Medicine (United States)*, 99, 2020.

[83] Wei Song, Hao Wang, and Qingyu Wu. Atrial natriuretic peptide in cardiovascular biology and disease (nppa). *Gene*, 569, 2015.

[84] Susana M.Chuva De Sousa Lopes, Rutger J. Hassink, Alie Feijen, Marga A. Van Rooijen, Pieter A. Doevendans, Leon Tertoolen, Aart Brutel De La Rivière, and Christine L. Mummery. Patterning the heart, a template for human cardiomyocyte development. *Developmental Dynamics*, 235, 2006.

[85] Xiao Ling Luo, Peng Zhang, Xiangyuan Liu, Shiqian Huang, Sen Le Rao, Qiurong Ding, and Huang Tian Yang. Myosin light chain 2 marks differentiating ventricular cardiomyocytes derived from human embryonic stem cells. *Pflugers Archiv European Journal of Physiology*, 473, 2021.

[86] Farah Sheikh, Robert C. Lyon, and Ju Chen. Functions of myosin light chain-2 (myl2) in cardiac muscle and disease. *Gene*, 569, 2015.

[87] Maria Gutierrez-Arcelus, Tuuli Lappalainen, Stephen B. Montgomery, Alfonso Buil, Halit Ongen, Alisa Yurovsky, Julien Bryois, Thomas Giger, Luciana Romano, Alexandra Planchon, Emilie Falconnet, Deborah Bielser, Maryline Gagnebin, Ismael Padioleau, Christelle Borel, Audrey Letourneau, Periklis Makrythanasis, Michel Guipponi, Corinne Gehrig, Stylianos E. Antonarakis, and Emmanouil T. Dermitzakis. Passive and active dna methylation and the interplay with genetic variation in gene regulation. *eLife*, 2013, 2013.

[88] Boxiang Liu, Milos Pjanic, Ting Wang, Trieu Nguyen, Michael Gloudemans, Abhiram Rao, Victor G. Castano, Sylvia Nurnberg, Daniel J. Rader, Susannah Elwyn, Erik Ingelsson, Stephen B. Montgomery, Clint L. Miller, and Thomas Quertermous. Genetic regulatory mechanisms of smooth muscle cells map to coronary artery disease risk loci. *American Journal of Human Genetics*, 103, 2018.

[89] Shao Ping Zhang, Ruo Han Yang, Jia Shang, Ting Gao, Rui Wang, Xiao Dong Peng, Xiao Miao, Lei Pan, Wen Jun Yuan, Li Lin, and Qi Kuan Hu. Foxc1 upregulates the expression of toll-like receptors in myocardial ischaemia. *Journal of Cellular and Molecular Medicine*, 23, 2019.

[90] Seungwoon Seo and Tsutomu Kume. Forkhead transcription factors, foxc1 and foxc2, are required for the morphogenesis of the cardiac outflow tract. *Developmental Biology*, 296, 2006.

[91] Nicole Wagner and Kay Dietrich Wagner. Every beat you take—the wilms tumor suppressor wt1 and the heart. *International Journal of Molecular Sciences*, 22, 2021.

[92] Raja Chakraborty, Fatima Zahra Saddouk, Ana Catarina Carrao, Diane S. Krause, Daniel M. Greif, and Kathleen A. Martin. Promoters to study vascular smooth muscle. *Arteriosclerosis, thrombosis, and vascular biology*, 39, 2019.

[93] Vitaly Sorokin, Keeran Vickneson, Theo Kofidis, Chin Cheng Woo, Xiao Yun Lin, Roger Foo, and Catherine M. Shanahan. Role of vascular smooth muscle cell plasticity and interactions in vessel wall inflammation. *Frontiers in Immunology*, 11, 2020.

[94] Jian Ruan, Lu Zhang, Donghua Hu, Xianghu Qu, Fan Yang, Fuxue Chen, Xiangqin He, Jian Shen, Kunzhe Dong, Megan Sweet, Christina Sanchez, Deqiang Li, Weinian Shou, Jiliang Zhou, and Chen Leng Cai. Novel myh11 dual reporter mouse model provides definitive labeling and identification of smooth muscle cells - brief report. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 2021.

[95] Ivan Zanoni and Francesca Granucci. Role of cd14 in host protection against infections and in metabolism regulation. *Frontiers in Cellular and Infection Microbiology*, 4, 2013.

[96] Malgorzata Burek, Paula A. Arias-Loza, Norbert Roewer, and Carola Y. Förster.

Claudin-5 as a novel estrogen target in vascular endothelium. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 30, 2010.

[97] Monika Litviňuková, Carlos Talavera-López, Henrike Maatz, Daniel Reichart, Catherine L. Worth, Eric L. Lindberg, Masatoshi Kanda, Krzysztof Polanski, Matthias Heinig, Michael Lee, Emily R. Nadelmann, Kenny Roberts, Liz Tuck, Eirini S. Fasouli, Daniel M. DeLaughter, Barbara McDonough, Hiroko Wakimoto, Joshua M. Gorham, Sara Samari, Krishnaa T. Mahbubani, Kourosh Saeb-Parsy, Giannino Patone, Joseph J. Boyle, Hongbo Zhang, Hao Zhang, Anissa Viveiros, Gavin Y. Oudit, Omer Ali Bayraktar, J. G. Seidman, Christine E. Seidman, Michela Noseda, Norbert Hubner, and Sarah A. Teichmann. Cells of the adult human heart. *Nature*, 588, 2020.

[98] Joyce C.K. Man, Karel Van Duijvenboden, Peter H.L. Krijger, Ingeborg B. Hooijkaas, Ingeborg Van Der Made, Corrie De Gier-De Vries, Vincent Wakker, Esther E. Creemers, Wouter De Laat, Bastiaan J. Boukens, and Vincent M. Christoffels. Genetic dissection of a super enhancer controlling the nppa-nppb cluster in the heart. *Circulation Research*, 2021.

[99] J. D. Steimle and I. P. Moskowitz. Tbx5: A key regulator of heart development, 2017.

[100] Franziska S. Rathjens, Alica Blenkle, Lavanya M. Iyer, Anke Renger, Fahima Syeda, Claudia Noack, Andreas Jungmann, Matthias Dewenter, Karl Toischer, Ali El-Armouche, Oliver J. Müller, Larissa Fabritz, Wolfram Hubertus Zimmermann,

Laura C. Zelarayan, and Maria Patapia Zafeiriou. Preclinical evidence for the therapeutic value of tbx5 normalization in arrhythmia control. *Cardiovascular Research*, 117, 2021.

[101] Hilma Holm, Daniel F. Gudbjartsson, David O. Arnar, Gudmar Thorleifsson, Gudmundur Thorgeirsson, Hrafnhildur Stefansdottir, Sigurjon A. Gudjonsson, Aslaug Jonasdottir, Ellisiv B. Mathiesen, Inger Njølstad, Audhild Nyrnes, Tom Wilsgaard, Erin M. Hald, Kristian Hveem, Camilla Stoltenberg, Maja Lisa Løchen, Augustine Kong, Unnur Thorsteinsdottir, and Kari Stefansson. Several common variants modulate heart rate, pr interval and qrs duration. *Nature Genetics*, 42, 2010.

[102] T. K. Ghosh, E. A. Packham, A. J. Bonser, T. E. Robinson, S. J. Cross, and J. D. Brook. Characterization of the tbx5 binding site and analysis of mutations that cause holt-oram syndrome. *Human Molecular Genetics*, 10, 2001.

[103] Alex V. Postma, Judith B.A. Van De Meerakker, Inge B. Mathijssen, Phil Barnett, Vincent M. Christoffels, Aho Ilgun, Jan Lam, Arthur A.M. Wilde, Ronald H.Lekanne Deprez, and Antoon F.M. Moorman. A gain-of-function tbx5 mutation is associated with atypical holt-oram syndrome and paroxysmal atrial fibrillation. *Circulation Research*, 102, 2008.

[104] Wei Feng Jiang, Ying Jia Xu, Cui Mei Zhao, Xin Hua Wang, Xing Biao Qiu, Xu Liu, Shao Hui Wu, and Yi Qing Yang. A novel tbx5 mutation predisposes

to familial cardiac septal defects and atrial fibrillation as well as bicuspid aortic valve. *Genetics and Molecular Biology*, 43, 2020.

[105] Joyce Man, Phil Barnett, and Vincent M. Christoffels. Structure and function of the nppa–nppb cluster locus during heart development and disease. *Cellular and Molecular Life Sciences*, 75, 2018.

[106] Jeanne L. Theis, Michael T. Zimmermann, Jared M. Evans, Bruce W. Eckloff, Eric D. Wieben, Muhammad Y. Qureshi, Patrick W. O'Leary, and Timothy M. Olson. Recessive myh6 mutations in hypoplastic left heart with reduced ejection fraction. *Circulation: Cardiovascular Genetics*, 8, 2015.

[107] Maximilian G. Posch, Stephan Waldmuller, Melanie Müller, Thomas Scheffold, David Fournier, Miguel A. Andrade-Navarro, Bernard de Geeter, Sophie Guillaumont, Claire Dauphin, Dany Yousseff, Katharina R. Schmitt, Andreas Perrot, Felix Berger, Roland Hetzer, Patrice Bouvagnet, and Cemil Özcelik. Cardiac alpha-myosin (myh6) is the predominant sarcomeric disease gene for familial atrial septal defects. *PLoS ONE*, 6, 2011.

[108] -myosin heavy chain: A sarcomeric gene associated with dilated and hypertrophic phenotypes of cardiomyopathy. *Circulation*, 112, 2005.

[109] Discovering motifs in ranked lists of dna sequences. *PLoS Computational Biology*, 3, 2007.

[110] Eran Eden, Roy Navon, Israel Steinfeld, Doron Lipson, and Zohar Yakhini. Go-

rilla: A tool for discovery and visualization of enriched go terms in ranked gene lists. *BMC Bioinformatics*, 10, 2009.

[111] Andrew E. Teschendorff and Ning Wang. Improved detection of tumor suppressor events in single-cell rna-seq data. *npj Genomic Medicine*, 5, 2020.

[112] Ivan P.G. Moskowitz, Anne Pizard, Vickas V. Patel, Benoit G. Bruneau, Jae B. Kim, Sabina Kupershmidt, Dan Roden, Charles I. Berul, Christine E. Seidman, and Jonathan G. Seidman. The t-box transcription factor tbx5 is required for the patterning and maturation of the murine cardiac conduction system. *Development*, 131, 2004.

[113] David E. Arnolds, Fang Liu, John P. Fahrenbach, Gene H. Kim, Kurt J. Schillinger, Scott Smemo, Elizabeth M. McNally, Marcelo A. Nobrega, Vickas V. Patel, and Ivan P. Moskowitz. Tbx5 drives scn5a expression to regulate cardiac conduction system function. *Journal of Clinical Investigation*, 122, 2012.

[114] Beth A. Firulli, Rajani M. George, Jade Harkin, Kevin P. Toolan, Hongyu Gao, Yunlong Liu, Wenjun Zhang, Loren J. Field, Ying Liu, Weinian Shou, Ronald Mark Payne, Michael Rubart-Von Der Lohe, and Anthony B. Firulli. Hand1 loss-of-function within the embryonic myocardium reveals survivable congenital cardiac defects and adult heart failure. *Cardiovascular Research*, 116, 2020.

[115] Asha Acharya, Seung Tae Baek, Guo Huang, Banu Eskiocak, Sean Goetsch, Caro-

line Y. Sung, Serena Banfi, Marion F. Sauer, Gregory S. Olsen, Jeremy S. Duffield, Eric N. Olson, and Michelle D. Tallquist. The bhlh transcription factor tcf21 is required for lineagespecific emt of cardiac fibroblast progenitors. *Development (Cambridge)*, 139, 2012.

[116] Bin Wang, Joel Weidenfeld, Min Min Lu, Shanna Maika, William A. Kuziel, Edward E. Morrisey, and Philip W. Tucker. Foxp1 regulates cardiac outflow tract, endocardial cushion morphogenesis and myocyte proliferation and maturation. *Development*, 131, 2004.

[117] F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. Scanpy: Large-scale single-cell gene expression data analysis. *Genome Biology*, 19, 2018.

[118] Tiffany B. Duong, Padmapriyadarshini Ravisankar, Yuntao Charlie Song, Jacob T. Gafranek, Ariel B. Rydeen, Tracy E. Dohn, Lindsey A. Barske, J. Gage Crump, and Joshua S. Waxman. Nr2f1a balances atrial chamber and atrioventricular canal size via bmp signaling-independent and -dependent mechanisms. *Developmental Biology*, 434, 2018.

[119] Fahima Syeda, Paulus Kirchhof, and Larissa Fabritz. Pitx2-dependent gene regulation in atrial fibrillation and rhythm control. *Journal of Physiology*, 595, 2017.

[120] Juan Wang, Pradhan Abhinav, Yingjia Xu, Ruogu Li, Min Zhang, Xingbiao Qiu, Ruomin Di, Qi Qiao, Xiumei Li, Ritai Huang, Song Xue, and Yiqing Yang. Nr2f2

loss of function mutation is responsible for congenital bicuspid aortic valve. *International Journal of Molecular Medicine*, 43, 2019.

[121] Catherine A. Risebro, Richelle G. Searles, Athalie A.D. Melville, Elisabeth Ehler, Nipurna Jina, Sonia Shah, Jacky Pallas, Mike Hubank, Miriam Dillard, Natasha L. Harvey, Robert J. Schwartz, Kenneth R. Chien, Guillermo Oliver, and Paul R. Riley. Prox1 maintains muscle structure and growth in the developing heart. *Development*, 136, 2009.

[122] Jasna Medvedovic, Anja Ebert, Hiromi Tagoh, and Meinrad Busslinger. Pax5: A master regulator of b cell development and leukemogenesis, 2011.

[123] Christrian Steinberg, Katharina Eisenächer, Olaf Gross, Wolfgang Reindl, Frank Schmitz, Jürgen Ruland, and Anne Krug. The ifn regulatory factor 7-dependent type i ifn response is not essential for early resistance against murine cytomegalovirus infection. *European Journal of Immunology*, 39, 2009.

[124] Vanessa Sue Wacleche, Aelie Cattin, Jean Philippe Goulet, Dominique Gauchat, Annie Gosselin, Auŕelie Cleret-Buhot, Yuwei Zhang, Ćecile L. Tremblay, Jean Pierre Routy, and Petronela Ancuta. Cd16+ monocytes give rise to cd103+raldh2+tcf4+ dendritic cells with unique transcriptional and immunological features. *Blood Advances*, 2, 2018.

[125] S. L. Nutt, D. Eberhard, M. Horcher, A. G. Rolink, and M. Busslinger. Pax5

determines the identity of b cells from the beginning to the end of b-lymphopoiesis. *International Reviews of Immunology*, 20, 2001.

[126] Chao Yang, Jason R. Siebert, Robert Burns, Zachary J. Gerbec, Benedetta Bonacci, Amy Rymaszewski, Mary Rau, Matthew J. Riese, Sridhar Rao, Karen Sue Carlson, John M. Routes, James W. Verbsky, Monica S. Thakar, and Subramaniam Malarkannan. Heterogeneity of human bone marrow and blood natural killer cells defined by single-cell transcriptome. *Nature Communications*, 10, 2019.

[127] Lisa Kalman, Mary Lou Lindegren, Lisa Kobrynski, Robert Vogt, Harry Hannon, Joelyn Tonkin Howard, and Rebecca Buckley. Mutations in genes required for t-cell development: Il7r, cd45, il2rg, jak3, img1, img2, artemis, and ada and severe combined immunodeficiency: Huge review. *Genetics in Medicine*, 6, 2004.

[128] H. M. Massey and Adrienne M. Flanagan. Human osteoclasts derive from cd14-positive monocytes. *British Journal of Haematology*, 106, 1999.

[129] T. William Mudd, Chunwan Lu, John D. Klement, and Kebin Liu. Ms4a1 expression and function in t cells in the colorectal cancer tumor microenvironment. *Cellular Immunology*, 360, 2021.

[130] Lyudmyla Kompaniyets, Alyson B. Goodman, Brook Belay, David S. Freedman, Marissa S. Sucosky, Samantha J. Lange, Adi V. Gundlapalli, Tegan K. Boehmer, and Heidi M. Blanck. Body mass index and risk for covid-19–related hospitaliza-

tion, intensive care unit admission, invasive mechanical ventilation, and death — united states, march–december 2020. *MMWR Surveillance Summaries*, 70, 2021.

[131] David R. Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16, 2004.

[132] Fast and accurate clustering of single cell epigenomes reveals cis-regulatory elements in rare cell types. *bioRxiv*, 2019.

[133] The encyclopedia of dna elements (encode): Data portal update. *Nucleic Acids Research*, 46, 2018.

[134] Hemant Suryawanshi, Robert Clancy, Pavel Morozov, Marc K. Halushka, Jill P. Buyon, and Thomas Tuschl. Cell atlas of the foetal human heart and implications for autoimmune-mediated congenital heart block. *Cardiovascular Research*, 116, 2020.

[135] Guang Li, Adele Xu, Sopheak Sim, James R. Priest, Xueying Tian, Tooba Khan, Thomas Quertermous, Bin Zhou, Philip S. Tsao, Stephen R. Quake, and Sean M. Wu. Transcriptomic profiling maps anatomically patterned subpopulations among single embryonic cardiac cells. *Developmental Cell*, 39, 2016.

[136] Mélanie Lambert, Samy Jambon, Sabine Depauw, and Marie Hélène David-Cordonnier. Targeting transcription factors for cancer treatment, 2018.

[137] Daniel M. DeLaughter, Alexander G. Bick, Hiroko Wakimoto, David McKean,

Joshua M. Gorham, Irfan S. Kathiriya, John T. Hinson, Jason Homsy, Jesse Gray, William Pu, Benoit G. Bruneau, J. G. Seidman, and Christine E. Seidman. Single-cell resolution of temporal gene expression during heart development. *Developmental Cell*, 39, 2016.

[138] Tahmina Samad and Sean M. Wu. Single cell rna sequencing approaches to cardiac development and congenital heart disease. *Seminars in Cell and Developmental Biology*, 118, 2021.

[139] V. A. Traag, L. Waltman, and N. J. van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, 9, 2019.

[140] Etienne Becht, Leland McInnes, John Healy, Charles Antoine Dutertre, Immanuel W.H. Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W. Newell. Dimensionality reduction for visualizing single-cell data using umap. *Nature Biotechnology*, 37, 2019.

[141] Bram Van de Sande, Christopher Flerin, Kristofer Davie, Maxime De Waegeneer, Gert Hulselmans, Sara Aibar, Ruth Seurinck, Wouter Saelens, Robrecht Cannoodt, Quentin Rouchon, Toni Verbeiren, Dries De Maeyer, Joke Reumers, Yvan Saeys, and Stein Aerts. A scalable scenic workflow for single-cell gene regulatory network analysis. *Nature Protocols*, 15, 2020.

[142] Cathy J. Hatcher, Marsha M. Goldstein, Caroline S. Mah, C. Susan Delia, and

Craig T. Basson. Identification and localization of tbx5 transcription factor during human cardiac morphogenesis. *Developmental Dynamics*, 219, 2000.

[143] C T Basson, D R Bachinsky, R C Lin, T Levi, J A Elkins, J Soults, D Grayzel, E Kroumpouzou, T A Traill, J Leblanc-Straceski, B Renault, R Kucherlapati, J G Seidman, and C E Seidman. Mutations in human tbx5 [corrected] cause limb and cardiac malformation in holt-oram syndrome [published erratum appears in nat genet 1997 apr;15(4):411]. *Nature Genetics*, 15, 1997.

[144] Frederick P. Roth, Jason D. Hughes, Preston W. Estep, and George M. Church. Finding dna regulatory motifs within unaligned noncoding sequences clustered by whole-genome mrna quantitation. *Nature Biotechnology*, 16, 1998.

[145] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck, Shiwei Zheng, Andrew Butler, Maddie J. Lee, Aaron J. Wilk, Charlotte Darby, Michael Zager, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P. Mimitou, Jaison Jain, Avi Srivastava, Tim Stuart, Lamar M. Fleming, Bertrand Yeung, Angela J. Rogers, Juliana M. McElrath, Catherine A. Blish, Raphael Gottardo, Peter Smibert, and Rahul Satija. Integrated analysis of multimodal single-cell data. *Cell*, 184, 2021.

[146] Zhana Duren, Xi Chen, Mahdi Zamanighomi, Wanwen Zeng, Ansuman T. Satpathy, Howard Y. Chang, Yong Wang, and Wing Hung Wong. Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations. *Pro-*

ceedings of the National Academy of Sciences of the United States of America, 115, 2018.

[147] Panwen Wang, Jing Qin, Yiming Qin, Yun Zhu, Lily Yan Wang, Mulin Jun Li, Michael Q. Zhang, and Junwen Wang. Chip-array 2: Integrating multiple omics data to construct gene regulatory networks. *Nucleic Acids Research*, 43, 2015.

[148] Adriana C. Gittenberger-De Groot, Margot M. Bartelings, Marco C. Deruiter, and Robert E. Poelmann. Basics of cardiac development for the understanding of congenital heart malformations. *Pediatric Research*, 57, 2005.

[149] Hana Imrichová, Gert Hulselmans, Zeynep Kalender Atak, Delphine Potier, and Stein Aerts. I-cistarget 2015 update: Generalized cis-regulatory enrichment analysis in human, mouse and fly. *Nucleic Acids Research*, 43, 2015.

[150] Joonhee Han, Ji Xiong, Dong Wang, and Xiang Dong Fu. Pre-mrna splicing: Where and when in the nucleus. *Trends in Cell Biology*, 21, 2011.

[151] Valentina Boeva. Analysis of genomic sequence motifs for deciphering transcription factor binding and transcriptional regulation in eukaryotic cells. *Frontiers in Genetics*, 7, 2016.

[152] Bum Kyu Lee, Akshay A. Bhinge, Anna Battenhouse, Ryan M. McDaniell, Zheng Liu, Lingyun Song, Yunyun Ni, Ewan Birney, Jason D. Lieb, Terrence S. Furey, Gregory E. Crawford, and Vishwanath R. Iyer. Cell-type specific and combinato-

rial usage of diverse transcription factors revealed by genome-wide binding studies in multiple human cells. *Genome Research*, 22, 2012.