Development and application of combinatorial single cell methods to tissue physiology and disease

David F Read

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2022

Reading Committee:

Cole Trapnell, Chair

William Noble

Doug Fowler

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

Abstract

Development and application of combinatorial single cell methods to tissue physiology and disease

David F Read

Chair of the Supervisory Committee:

Cole Trapnell

Department of Genome Sciences

Advances in single-cell sequencing technologies present an opportunity to study the cellular and molecular basis of complex tissue physiology at global scale with unprecedented throughput. In this work, methods for application of a method of single cell sequencing – single-cell combinatorial indexing ("sci") methods – were extended for use in mammalian tissue. In the human heart, analysis of single nucleus data reveals variation by age and sex in healthy donors, divergent use of transcription factor motifs in adult vs. fetal chromatin, and distal genomic sites that improve predictive models of cell type-specific expression. In mouse models of lung disease, we find aberrant differentiation states in the macrophages of a model of pulmonary alveolar proteinosis, emergence of an osteoclast-like macrophage phenotype in a model of silicosis, and in both models we find widespread alterations across the cell types of the lung. In

overlapping work, we build interpretable predictive models of gene expression in the human heart and in the *P. falciparum* parasite, finding cell-type specific transcription factors in adult human heart and uncovering strikingly stage-specific information content in histone marks in *Plasmodium*.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Science is a team sport. Both within the lab and in life at large, my experience has been that research is deeply interwoven with the people around us.

Within the University of Washington, I owe a great thanks to my lab mates who have been the closest to me in the daily, monthly, and yearly slogs of graduate school. It's been a privilege to get to work with such a talented group of scientists, particularly ones that are so generous with their advice and time. I also want to say "thank you" to my PhD cohort, whose camaraderie has been invaluable in navigating the twists and turns of the program. Finally, I also am grateful for the mentorship of my adviser, Cole Trapnell, who has been a continuous source of insights and guidance from the minute details of projects to the wide-ranging conversations on the future of biology.

In addition, getting through a PhD unscathed has been in no small part due to the tremendous support I received from outside of the lab. My family has been an unwavering pillar of support, whatever happened in my PhD and however cranky I felt towards science on a particular day. Friends in Seattle and around the country have likewise been steadfast in belief in my ability to push through any challenges and available to listen to my rambling accounts of graduate school. My girlfriend Rupali has been a constant source of affection, wisdom, and calm who has always helped me see the best in my present and future.

To all those who helped me through this degree: I could not have done it without you.

Chapter 1. INTRODUCTION

The work described in this dissertation largely comes from my work in the lab of Cole Trapnell, where I applied "sci" single-nucleus RNA-Seq methods to study healthy human heart biology and hallmarks of disease models in mouse lung. An additional significant area of work was done during and after my time as a rotation student in the lab of Bill Noble, where I led analysis and submission of a project using predictive models of gene expression to study *P. falciparum.*

Large portions of this dissertation are adapted with minimal modification from manuscripts that are either published (Chapter 2), under review/available as preprints (Chapter 3), or under preparation for submission (Chapter 4, separate papers for work on silicosis and pulmonary alveolar proteinosis). Any cases of such adaptation are noted immediately before the relevant section.

1.1 SINGLE CELL SEQUENCING

The cell is the fundamental unit of life. Even in contexts involving billions of cells – such as understanding physiology of complex tissues or organisms – an understanding of the processes carried out within specialized cells and interactions between cells is a cornerstone of bottom-up biological understanding.

Technologies for studying biology inherently make measurements at varying levels of cellular resolution. Microscopy techniques have long provided information at cellular or sub-cellular resolution. Methods such as fluorescence-activated cell sorting (FACS) similarly give cell-by-cell information (Herzenberg et al. 2002). In contrast, widely used high-throughput methods like RNA-Seq take measurements that inherently average over many cells within a

sample (Stark, Grzelak, and Hadfield 2019). Recent years have seen advances in single-cell technologies that combine genome- or transcriptome-wide measurements with single-cell resolution. Due to inherent biochemical attributes of the species to be measured, these methods are mostly developed for measurements of genetic (Fan et al. 2021; Gawad, Koh, and Quake 2016), epigenetic (Fang et al. 2021; Cusanovich et al. 2015; Satpathy et al. 2019), and transcriptomic (Zheng et al. 2017; Cao et al. 2017; Jaitin et al. 2014; Grün et al. 2015) data, whereas methods for measurements of proteins (Schoof et al. 2021) or metabolites (Seydel 2021) are in earlier stages of development.

Single-cell RNA-Seq methods capturing transcriptome-wide information in single cells are among the most developed single-cell methods. Published methods include plate-based approaches (Picelli et al. 2014), microfluidics (Zheng et al. 2017), and combinatorial indexing (Cao et al. 2017). Split-pool methods use combinatorial indexing to identify individual cells of origin for captured cDNA molecules, providing the potential to exponentially scale the number of transcriptomes captured as the number of indexing rounds increases (Cao et al. 2019). One method – sci RNA-Seq – offers the potential to sequence tens of thousands to millions of cells in a single experiment, providing improved statistical power, ability to detect rare cell types, and the potential to analyze larger numbers of unique samples than would be financially feasible with more expensive approaches.

My work in the Trapnell lab included a significant effort to develop methods related to split-pool "sci" RNA-Seq methods (Chapter 4, Section 4.1). The most fruitful result of that effort was a method for isolating and fixing nuclei from whole organs such that the resulting nuclei are compatible with 2- or 3-level sci RNA-Seq workflows. At the time, the method was – to the best of my knowledge – the first workflow to reproducibly generate sci RNA-Seq data from adult

mouse non-brain tissue and human heart. That method was used to generate datasets that were core elements of three manuscripts: one covering healthy human heart (Chapter Three), one exploring a mouse model of pulmonary alveolar proteinosis ("PAP", Chapter Four), and a third studying a mouse model of silicosis (Chapter Four).

1.2 PREDICTIVE MODELS OF GENE EXPRESSION

Transcription of RNA is regulated through concerted transcription factor (TF) activity, chromatin remodeling, and epigenetic modification (Casamassimi and Ciccodicola 2019). A standing challenge in the field of gene regulation is to understand the relationship between different aspects of gene regulation – such as presence of TF DNA motifs, covalent histone modifications, DNA accessibility, and so on – and gene expression. Eventually, the goal of such efforts is a causative understanding of the exact mechanisms underpinning gene expression, silencing, and control. However, defining causal links will inherently require significant work in perturbational experiments linking aspects of genomic biochemistry to transcription. As an intermediate step, a more immediately tractable goal is understanding the correlative relationships between sequence or epigenetic features and gene expression. Abundant assembled genomes (Rhie et al. 2021) and datasets covering epigenetic and TF-binding profiles in humans (ENCODE Project Consortium 2012), model organisms (Muers 2011), and non-model organisms (Bonasio 2015; Aurrecoechea et al. 2009) have been generated in recent years. With many such datasets in hand, significant work has already been undertaken to find the correlative relationships between sequence, epigenetic state, and transcription (Cheng et al. 2011; Dong et al. 2012; Pe'er, Regev, and Tanay 2002; Singh et al. 2016; Beer and Tavazoie 2004; Y. Chen et al. 2016; Agarwal and Shendure 2020; J. Zhou et al. 2018).

A widely used strategy for understanding the relationship between genomic or epigenomic features and RNA transcription is fitting predictive models. Gene expression – either as a binary high/low level in a classification task or as a numerical value in a regression framework – is modeled using sequence and epigenetic features in a machine learning model, with the trained models evaluated to understand the relative use of each feature in predicting gene expression (Cheng et al. 2011; Dong et al. 2012; Y. Y. Lu et al. 2018; Lundberg and Lee 2017; Narlikar et al. 2010). Such analyses contributed to understanding the association strength of chromatin accessibility with transcription (Cheng et al. 2011; Dong et al. 2012; Duren et al. 2017), the redundant information content of covalent histone modifications and TF binding profiles in simple models of transcription (McLeay et al. 2012), and the magnitude of regulatory predictive information in distal sites (Pliner et al. 2018).

My dissertation describes two contexts in which I used predictive models to understand molecular correlates of expression. The most extensive work involved fitting models in the malaria-causing parasite *Plasmodium falciparum* (Chapter 2). In addition, I developed a linear model to predict expression levels across cell types in the human heart (Chapter 3.3.5).

The following is adapted with minimal modifications from (Read et al. 2019)

Chapter 2. PREDICTING GENE EXPRESSION IN THE HUMAN MALARIA PARASITE PLASMODIUM FALCIPARUM USING HISTONE MODIFICATION, NUCLEOSOME POSITIONING, AND 3D LOCALIZATION FEATURES

2.1 ABSTRACT

Empirical evidence suggests that the malaria parasite *Plasmodium falciparum* employs a broad range of mechanisms to regulate gene transcription throughout the organism's complex life cycle. To better understand this regulatory machinery, we assembled a rich collection of genomic and epigenomic data sets, including information about transcription factor (TF) binding motifs, patterns of covalent histone modifications, nucleosome occupancy, GC content, and global 3D genome architecture. We used these data to train machine learning models to discriminate between high-expression and low-expression genes, focusing on three distinct stages of the red blood cell phase of the *Plasmodium* life cycle. Our results highlight the importance of histone modifications and 3D chromatin architecture in *Plasmodium* transcriptional regulation and suggest that AP2 transcription factors may play a limited regulatory role, perhaps operating in conjunction with epigenetic factors.

2.2 INTRODUCTION

Plasmodium falciparum is the deadliest species of malaria parasite, responsible for 445,000 deaths in 2016 (Programme 2017). As resistance to antimalarial drugs spreads, demand

for novel antimalarials increases. Designing such novel drugs would require an improved understanding of the biology of this parasite. Currently, one of the primary open questions in *Plasmodium* biology is how the parasite maintains precise control of gene expression. The current work aims to address this question by constructing an accurate predictive model of *P. falciparum* transcription. The model accounts for the rich landscape of transcriptional control mechanisms in *Plasmodium* by incorporating five types of features, representing transcription factor (TF) binding, covalent histone modifications, nucleosome positioning, GC content, and chromatin 3D structure.

In many eukaryotes, TF binding within and around gene promoters is considered the dominant mechanism of gene expression control. However, in *Plasmodium*, several lines of evidence suggest that TF binding may be less central to transcriptional control. First, although major components of the general transcription machinery are present in the *Plasmodium* genome (Coulson, Hall, and Ouzounis 2004), a relatively small set of specific *Plasmodium* TFs (~27) have been identified and validated in the parasite genome(Coulson, Hall, and Ouzounis 2004). In comparison, the similarly sized genome of the yeast *S. cerevisiae* contains ~170 specific TFs (de Boer and Hughes 2012). Second, among the subset of TFs whose binding affinities have been characterized via *in vitro* protein binding microarrays (De Silva et al. 2008), only a handful display stage-restricted expression and play clear roles in regulating life cycle transitions. An example is PfAP2-G, which drives expression of gametocyte-specific genes (Kafsack et al. 2014). Third, a large number of *Plasmodium* genes are predicted by homology to function in the regulation of chromatin structure, mRNA decay, and translation (Coulson, Hall, and Ouzounis 2004), suggesting the importance of epigenetic and post-transcriptional regulation.

Among mechanisms for epigenetic regulation, patterns of covalent histone modifications are perhaps the most widely studied and understood. In this respect, some aspects of *P. falciparum* gene regulation are shared with other eukaryotes, including the presence of the typically heterochromatin-associated H3K9me3 histone modification at repressed *var* genes (referred to as virulence genes, for their role in parasite pathogenicity) (Chookajorn et al. 2007) and depletion of promoter nucleosomes correlating with gene transcription (Bunnik et al. 2014). On the other hand, *Plasmodium* epigenomic dynamics also exhibit notable deviations from those in commonly studied eukaryotes, such as abundant and broad distributions of activating histone marks (Ay et al. 2015; Lopez-Rubio, Mancio-Silva, and Scherf 2009) or active histone variant H2Z in the promoters of all genes with the exception of genes involved in immune evasion (Petter et al. 2013), an absence of H3K27me3 repressive marks (Ay et al. 2015), and genome-wide changes in nucleosome occupancy during the asexual cycle (Bunnik et al. 2014; Ponts et al. 2010). These observations suggest that the parasite may make use of a ``histone code" like other well-characterized eukaryotes, though the specific role of individual elements may differ.

In addition, empirical evidence suggests that gene regulation in *Plasmodium* occurs through changes in chromatin structure, including shifts in nucleosome occupancy at the local level and 3D positioning at larger scales. Nucleosome occupancy, as measured by MNase-assisted isolation of nucleolar elements (MAINE) and formaldehyde-assisted isolation of regulatory elements (FAIRE), or assay for transposase accessible chromatin with high-throughput sequencing (ATAC-seq), exhibit cyclic patterns that closely track changes in gene expression during the red blood cell (erythryocytic) stages of the parasite life cycle (Ponts et al. 2010; Toenhake et al. 2018). In addition, 3D models of *Plasmodium* DNA based on Hi-C assays at multiple time points during the red blood cell (Ay et al. 2015) and transmission (Bunnik et al.

2018) stages of the parasite point to a ``gradient'' of expression across the nucleus, from a repressive center near the telomeres to an expressive center at the centromeres.

Based on the above evidence, we hypothesized that the cascade of transcripts observed throughout the red blood cell (erythrocytic) cycle is the result of a combination of transcription factor binding, histone modifications, and changes in chromatin structure. We further predicted than an integrated analysis of the relationships among transcription and TF binding, histone modification, and chromatin structure data could reveal the relative significance of individual features in defining high- and low-expression genes.

We are not the first to build predictive models of gene expression (Table 2.1), though to our knowledge we are the first to do so in *Plasmodium*. To keep the model simple, we focus on the binary classification task, in which each gene is either ``on'' or ``off,'' rather than the more challenging regression setting. Prediction of gene expression has been framed as a classification task in numerous previous works, with our approach most closely resembling the analysis in references (Cheng et al. 2011; Dong et al. 2012; Singh et al. 2016). Furthermore, because we sought to determine the importance of features within individual stages we restrict ourselves to predicting relative high- or low-expression labels with respect a single parasitic stage at a time, rather than developing a single model that predicts absolute expression values irrespective of stage. Accordingly, we build separate models in three different stages of the *P. falciparum* life cycle and analyze the resulting models to understand which features are implicated in the up- or down-expression of *Plasmodium* genes in different stages of the parasitic life cycle.

Model	Year	Yeast	Mouse	Human	mRNA expression	TF motifs	DNA sequence	TF ChIP-seq	Histone ChIP-seq	DNase or ATAC	PolII ChIP-seq	Classification	Regression
Multiple linear regression (Bussemaker, Li, and Siggia 2001)	2001	✓				✓	✓						✓
Conditional probability given levels of regulators (Pe'er, Regev, and Tanay 2002)	2002	✓			✓							✓	
Classification tree (Segal, Yelensky, and Koller 2003)	2003	✓			✓							✓	
Bayesian network (Beer and Tavazoie 2004)	2004	✓			✓		✓					✓	
Boosted alternating decision trees (Middendorf et al. 2004)	2004	✓			✓	✓						✓	
Boosted alternating decision trees (Kundaje et al. 2006)	2006	✓			✓	✓						✓	
Principal component regression and regression tree (Ouyang, Zhou, and Wong 2009)	2009		✓					✓					✓
Multiple linear regression (Karlić et al. 2010)	2010			✓					✓				✓
Support vector machine (Cheng et al. 2011)	2011		✓						✓			✓	✓
Random forest and multiple linear regression (Dong et al. 2012)	2012			✓					✓			✓	✓
Multiple log-linear regression (McLeay et al. 2012)	2012		✓					✓	✓	✓			✓
Bayesian variable selection regression (X. Zhou et al. 2014)	2014			✓					✓		✓		✓
Multiple regression (González, Setty, and Leslie 2015)	2015			✓		✓	✓		✓	✓			✓
Multilayer perceptron (Y. Chen et al. 2016)	2016			✓	✓								✓
Convolutional neural network (Singh et al. 2016)	2016			✓				✓				✓	
Multiple regression (Duren et al. 2017)	2017			✓	✓				✓	✓			✓
Convolutional neural network (Kelley et al. 2018)	2018			✓			✓						✓
Multitask regression (Osmanbeyoglu et al. 2019)	2019			✓		✓	✓			✓			✓

Table 2-1: **Comparison of methods for predicting gene expression.** The "mRNA Expression" feature involves using the mRNA expression of a set of putative regulators to predict the mRNA expression of a different set of target genes.

2.3 METHODS

2.3.1 *Datasets*

Feature	Study	Description	Time point		
			Ring	Trophozoite	Schizont
Distance to telomeres	(Ay et al. 2014)	Hi-C inferred distance	✓	✓	✓
Distance to centromeres	(Ay et al. 2014)	Hi-C inferred distance	✓	✓	✓
Distance to center	(Ay et al. 2014)	Hi-C inferred distance	✓	✓	✓
Nucleosome occupancy	(Bunnik et al. 2014)	100-200 base pair fragments	✓	✓	✓
H2A.z	(Bártfai et al. 2010)	ChIP-Seq	✓	✓	✓
H3K9Ac	(Bártfai et al. 2010)	ChIP-Seq	✓	✓	✓
H3K4me3 (Bartfai et al)	(Bártfai et al. 2010)	ChIP-Seq	✓	✓	✓
H3K36me2	(Jiang et al. 2013)	ChIP-Seq	✓		✓
H3K36me3	(Jiang et al. 2013)	ChIP-Seq	✓		✓
H3K9me3	(Jiang et al. 2013)	ChIP-Seq	✓		✓
H4K20me3	(Jiang et al. 2013)	ChIP-Seq	✓		✓
H3K4me3 (Jiang et al)	(Jiang et al. 2013)	ChIP-Seq	✓		✓
GC content	See Methods	100 bp sliding windows	N/A	N/A	N/A
TF motifs	(Weirauch et al. 2014)	FIMO scan of TF motifs	N/A	N/A	N/A

Table 2-2: **Summary of datasets used in classification models.** Dataset sources are shown, along with the time points from each study which were used to represent each of three life cycle stages. "N/A" is listed for the GC content and motif score features, as they did not vary across life cycle stages.

Although *P. falciparum* passes through multiple stages---mosquito, human liver, and human blood---we focus here on the human blood stage of the parasite life cycle, primarily because of the availability of a wide number of relevant data sets. We gathered data for three time points, corresponding to the three main asexual stages within the red blood cell cycle: ring,

trophozoite and schizont. Most data sets described below (Table 2.2) are available in all three of these time points, with the exception of some ChIP-seq data for covalent histone modifications (H3K36me2, H3K36me3, H3K9me3, H4K20me3, and one replicate of H3K4me3 (Jiang et al. 2013)) that were not available for the trophozoite stage.

2.3.2 *GRO-seq*

To define the on/off labels for our classifier, we used the GRO-seq values from Lu *et al.* (X. M. Lu et al. 2017). We used GRO-seq values normalized for GC content, gene length, and the "parasitemia factor" of a stage (X. M. Lu et al. 2017). To generate binary labels for genes, for each stage we sorted all protein-coding genes by the normalized GRO-seq counts assigned to that gene in that stage. We labeled the top third of genes as ``High expression" and the bottom third of genes as ``Low Expression" (Figure 2.1). The middle third of genes were not used in the analysis.

The decision to use tertiles rather than, say, quartiles or simply dividing at the median was somewhat arbitrary. Past works have used a number of schemes, such as dividing genes into tertiles (Kundaje et al. 2006; Middendorf et al. 2004), dividing genes in half at the median (Singh et al. 2016; Cheng et al. 2011), or dividing by zero/non-zero status (Dong et al. 2012). We elected to use tertiles and perform classification using the top and bottom sets in part to make the classification task somewhat easier (by only giving the model examples that are well-separated by expression) and in part to limit the detrimental effect of possible noise in the GRO-seq data (because noise is less likely to flip a gene between classes when the divisions are made at tertiles than if divisions were made at the median).

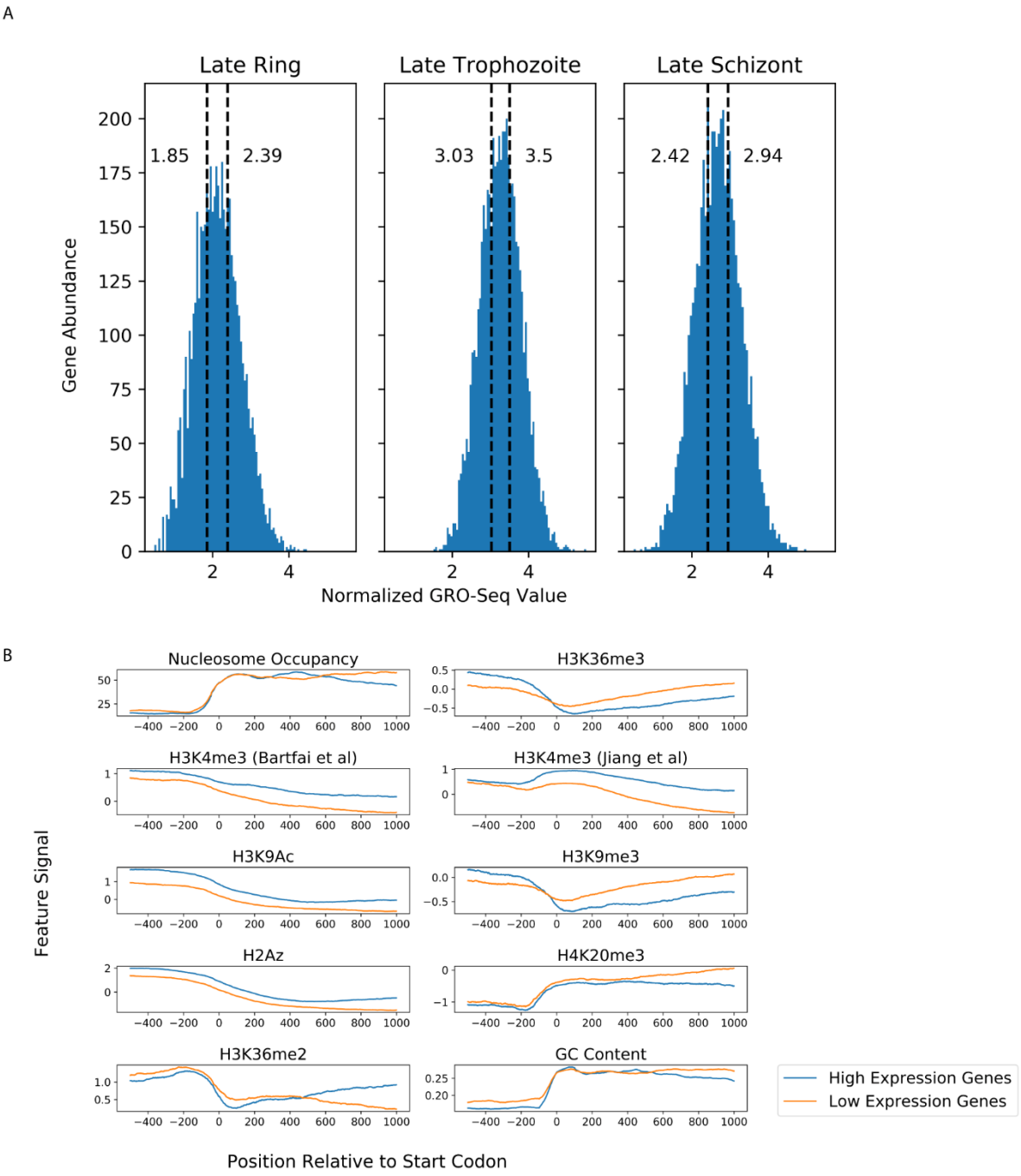

Figure 2-1: **Differences between high- and low-expression genes.** (A) Histograms showing the distribution of normalized GRO-Seq values assigned to protein-coding genes within each of three stages of the P. falciparum life cycle. The labeled, dashed vertical lines indicate the cut-off values for genes categorized as low- and high-expression. (B) Aggregation plots showing the average signal for features with respect to the start codons of high-expression (blue) and low-expression (orange) genes in the ring stage.

2.3.3 *Transcription start site and coding sequence annotations*

Many of the features that we employed require specifying the start coordinate of each given gene. For this purpose, we use two data sets of coordinates: either the start codons from the PlasmoDB v29 annotation or transcription start sites based on CAGE-Seq data from (Adjalley et al. 2016). In that resource, multiple start sites are often annotated for a given protein coding gene. To assign a single TSS for use in feature assignment, we first looked to see if the ``primary TSS" assigned in (Adjalley et al. 2016) was upstream of the start codon of a gene. If it was, then the start of that TSS was used. Otherwise, we used the TSS lying upstream and closest to the start codon. If no annotated TSS was upstream of the start codon of a gene, then the start codon was used.

2.3.4 *Histone modification ChIP-seq*

ChIP-seq data for the following histone modifications were collected from two studies: H3K4me3, H3K9Ac, and H2.Az from Bartfai et al. (Bártfai et al. 2010) and H3K36me2, H3K36me3, H3K9me3, H4K20me3, and H3K4me3 from Jiang et al. (Jiang et al. 2013). Note that one mark, H3K4me3, was measured in both studies. All of the ChIP-seq data was reanalyzed using a standard pipeline that consisted of trimming reads to 76 nucleotides using the fastx toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html), mapping reads to the *P. falciparum* genome (PlasmoDB v29) using bwa-mem (H. Li and Durbin 2010), filtering unmapped and multimapping reads using samtools (H. Li et al. 2009), and generating bedgraph files using bedtools (Quinlan and Hall 2010). Fold-change over background was calculated relative to input DNA where available. For histone ChIP-seq datasets from Jiang *et al.*, no input DNA data was available, so values were calculated relative to the mean signal over all the data.

2.3.5 *Nucleosome occupancy*

MNase data from (Bunnik et al. 2014) was downloaded in FASTQ format, then trimmed and filtered using sickle version 1.33 (https://github.com/najoshi/sickle). Reads were aligned to the *P. falciparum* genome (PlasmoDB v29) using bwa-mem (H. Li and Durbin 2010), then sorted and filtered for mapped reads using samtools (H. Li et al. 2009). A custom Python script selected all alignments between 100 and 200 bp in length, which were then used to generate a bedgraph file using genomeCoverageBed (Quinlan and Hall 2010).

2.3.6 *Hi-C*

Per-gene distances from telomere centroid, centromere centroid, and center were computed in a previous study carried out by our lab (Ay et al. 2015), using 3D models generated from Hi-C data using PASTIS (Varoquaux et al. 2014). These values were obtained directly from https://noble.gs.washington.edu/proj/plasmo3d/.

2.3.7 *DNA sequence features*

Position-frequency matrices for 25 AP2 family transcription factors for *P. falciparum* were downloaded from CIS-BP (Weirauch et al. 2014). We focused on available AP2 family motifs for two reasons. First, AP2 transcription factors have been widely speculated to play a key role in TF-mediated transcriptional regulation throughout the erythrocytic cycle due to variable expression, sequence-specific DNA binding, and presence of AP2 motifs upstream of genes whose expression varies throughout erythrocytic stages (Balaji et al. 2005; Campbell et al. 2010). Second, the DNA binding specificities of AP2 transcription factors were evaluated using a high-throughput *in-vitro* protein binding microarray and subjected to *in vivo* validation (Campbell et al. 2010), generating a motif set derived by a consistent, rigorous workflow.

With each motif, we scanned the *P. falciparum* genome using FIMO (Grant, Bailey, and Noble 2011) with a p-value threshold of 0.01. This fairly permissive cut-off was arbitrarily set, leaving more aggressive feature selection for downstream model training and evaluation. To ensure that the background model represented the unique sequence context of *P. falciparum*, we generated a background model from the *P. falciparum* genome with the MEME Suite command fasta-get-markov with Markov order 1 (Bailey et al. 2009). In addition, percent GC was calculated in 101 base windows centered at each position in the genome.

Features based on histone modifications, H2Az composition, nucleosome occupancy, and GC content were segregated into ``promoter" and ``gene body" features. The ``promoter" feature was the mean feature signal from -500 bases up to the start codon, whereas the ``gene body" feature was the mean feature signal from the start codon to 1 kb into the coding sequence.

2.3.8 *Models*

We used three types of models to classify *Plasmodium* expression and select predictive features. The first was logistic regression with elastic net regularization, using the scikit-learn implementation (sklearn.linear_model.SGDClassifier). The second was a tree model with gradient boosting, using the XGboost Python implementation (T. Chen and Guestrin 2016). The third was a multi-layer perceptron model, with two hidden layers, each containing the same number of nodes as the input layer. This model was implemented by DeepPINK (Y. Y. Lu et al. 2018), which is designed to achieve robust feature selection with a controlled error rate.

2.3.9 *Performance metric*

The performance of each model was evaluated in terms of receiver operator characteristic (ROC) curves. These plots show the rate of true positive classifications (on the y-axis, indicating

sensitivity) against the rate of false positive classifications (on the x-axis, indicating 1 - specificity). The area under the ROC curve (AUROC) quantifies the ability of the classifier to balance sensitivity (true positives) against specificity (avoiding false positives). An AUC value of 1 corresponds to perfect performance, whereas a value of 0.5 corresponds to random guessing.

For Figure 2.2A, the ROC curve for logistic regression classification was generated by combining the gene scores from the test sets in three separate folds of cross-validation. These scores were sorted together to generate the combined ROC curve shown.

2.3.10 *Train/Test Data Splitting*

We split the *P. falciparum* into five approximately equally sized (by gene count) folds by chromosome: fold 1 included chromosomes 1, 3, and 13; fold two included 2, 9, and 11; fold three included 7 and 14; fold four included 6, 8, and 10; and fold five included 4, 5, and 12. This split was done by calculating the number of genes-per-fold in a perfectly even split, then choosing the division of chromosomes whose totals had the smallest mean-squared error from this ideal value, across all possible permutations of chromosome-to-fold assignments. A Python script that tests all permutations of chromosome sets, selecting the division that minimizes the mean squared error, is available in the Github repository, dataPreProcessing/selectingDataFolds/divideGenomeIntoFiveSets.py.

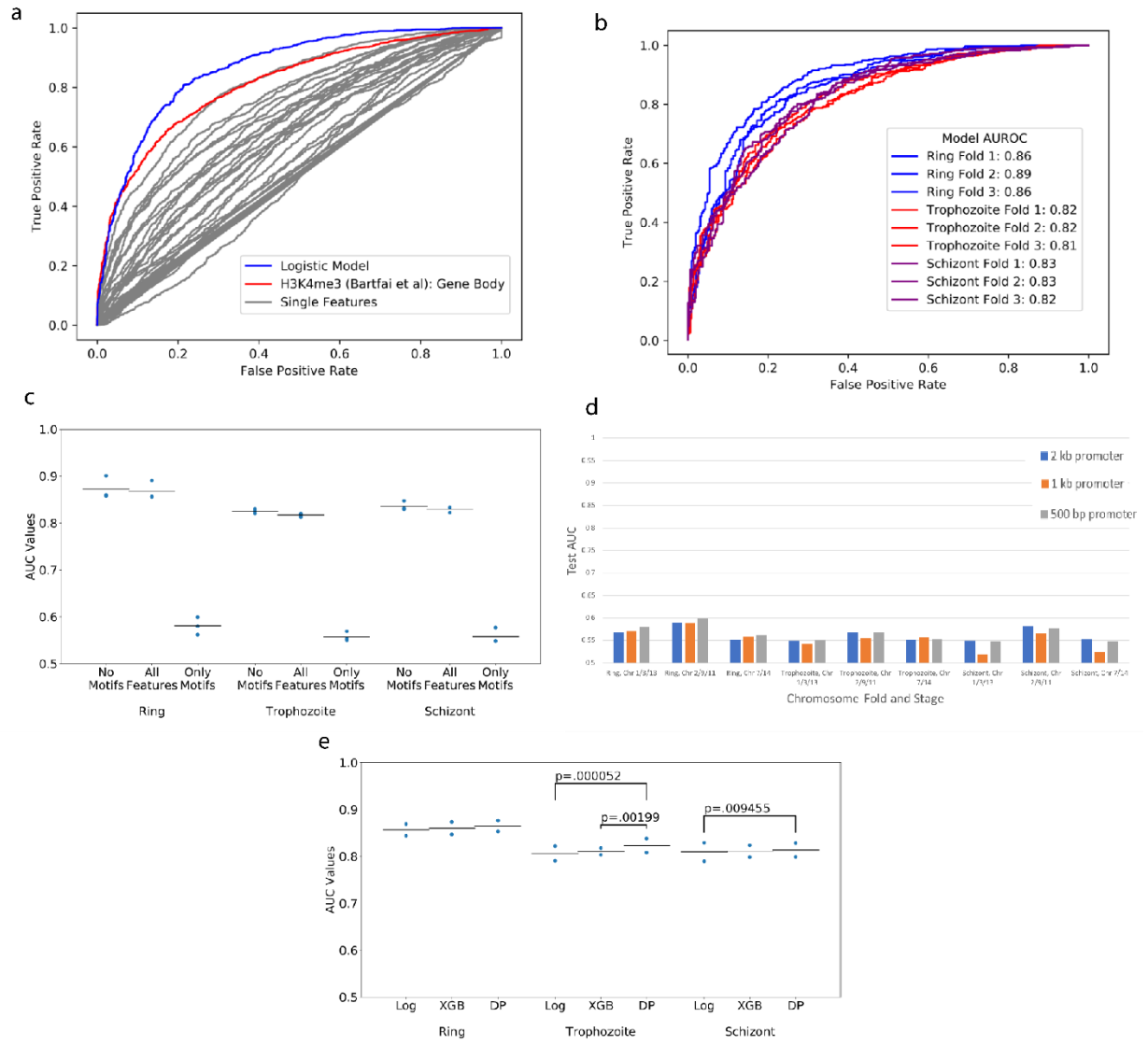

Figure 2-2: **Comparison of classification models.** A) Classification models outperform all individual features for use in classification of gene expression. Grey lines represent the ROC curves resulting from ranking genes by the values of single features in the ring stage, with the best-performing feature shown in red. The blue line represents the ROC curve from training a logistic regression model with elastic net regularization, where the curve is created by combining the predictions across all three test sets. B) The ROC curves for classification of gene expression by logistic regression across ring, trophozoite, and schizont stages. Individual curves represent performance in one of test three folds in cross-validation. C) AUROC values for logistic models trained with or without motif scores as features. Points represent AUROC values on the test set in three-fold cross-validation; bars represent average AUROC values on the test data. D) AUROC for models only using TF motif features, defined over varying promoter sizes. E) The AUROC values resulting from training of distinct models in different stages ("Log" = Logistic Regression, "XGB" = XGBoost, "DP" = DeepPINK). Individual points represent the AUROC values from distinct test sets, for the listed model in a given stage. Brackets are labeled with p values for pairwise comparisons within stages where $p < 0.05$, using the DeLong method for comparing AUC values.

2.3.11 *Model development and hyperparameter tuning*

The first three folds were used for feature development and hyperparameter tuning. During this stage, we selected hyperparameters by three-fold internal cross-validation. For the logistic regression model with elastic net regularization, we tuned the ``alpha" and ``l1_ratio" parameters in a sklearn.linear_model.SGDClassifier model.

Model	Parameter	Possible values	Selected value		
			Ring	Trophozoite	Schizont
LR	alpha	0.1, 0.01, 0.001, 0.0001	0.0001	0.0001	0.0001
LR	l1_ratio	0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95	0.95	0.8	0.9
Boosted trees	max_depth	4, 5, 6, 7, 8, 9	4	6	4
Boosted trees	min_child_weight	2, 3, 4, 5, 6, 7	5	5	6
Boosted trees	subsample	0.3, 0.4, 0.7, 0.8, 1.0	0.7	0.4	0.4
Boosted trees	colsample_bylevel	0.3, 0.5, 0.7, 1.0	0.3	0.5	0.3
Boosted trees	n_estimators	40, 60, 80, 100	100	60	80

Table 2-3: **Hyperparameter selection**

The ``alpha" value determines the magnitude of the regularization penalty relative to classification error, while ``l1_ratio"determines the relative magnitude of the L1 and L2 penalty terms (1 = pure LASSO penalty, 0 = purely ridge regularization). For the boosted trees model, we tuned the ``max_depth", ``min_child_weight", ``subsample", ``colsample_bylevel", and ``n_estimators"hyperparameters in an xgboost.XGBClassifier model. ``Max_depth"controls the tree depth of the decision trees composing the XGBoost ensemble, ``min_child_weight" controls the minimum weight in a leaf node that is allowed to be split further, ``subsample" controls the portion of training data samples for training each additional tree, ``colsample_by_level" controls whether re-sampling is done for each new depth level within trees, and ``n_estimators" is the number of trees in the model. In each case, we performed a grid search across the values listed in Table 2.3, testing all possible combinations of hyperparameter values using cross-validation within the three folds used for model development and selecting the hyperparameter combination with the lowest test error.

On the basis of initial analyses, we eliminated the motif-based features from our feature set, and we chose to use features based on CDS rather than TSS locations (see Section 2.4.3 for details).

2.3.12 *Model evaluation on test data*

Subsequently, we trained classifiers to make predictions on each of the two test folds, in each case training on the four remaining folds. In this case, hyperparameters were selected that yielded the greatest AUROC value in the three training set ``sub-folds" using cross-validation, as implemented in GridSearchCV in sklearn.grid_search. The selected hyperparameters are listed in Table 2.3. Probabilistic classification scores for all genes in both of the two test folds were combined for testing the statistical significance of differences in AUC values. AUC values were compared using the DeLong test for correlated AUCs (DeLong, DeLong, and Clarke-Pearson 1988) as implemented in the pROC package in the R language (Robin et al. 2011).

2.3.13 *XGBoost and SHAP Values*

The gradient boosting method XGBoost is powerful but challenging to interpret. XGBoost assigns classification labels by taking a consensus decision from an ensemble of individual decision trees (T. Chen and Guestrin 2016). XGBoost models are appealing due to their ability to capture complex interactions among features as well as non-linear relationships between features and classification labels (T. Chen and Guestrin 2016). However, understanding the importance of individual features within such ensembles is challenging, because the model may use a given feature in multiple locations across the individual trees (in contrast to a regression model with a readily interpretable coefficient assigned to a feature).

Consequently, we used SHAP (Lundberg and Lee 2017) to help interpret the trained XGBoost models. SHAP is a software package that quantifies the effect of each feature on the classification of each example (each gene, in our case) by measuring how much information that feature provides in addition to various subsets of other features being used in the model. The method obeys key mathematical properties and matches human intuition in tested cases (Lundberg and Lee 2017). Running SHAP on our trained XGBoost models provided us with ``SHAP values" for each feature, for every gene. These scores can be studied on a gene-by-gene basis and can be aggregated across all genes within a stage to obtain a consensus score, comparable to a regression coefficient.

2.3.14 *DeepPINK*

Similar to XGBoost, DeepPINK can also capture non-linear relationships between features and classification labels. Rather than boosted gradients, DeepPINK uses a deep neural network model. Importantly, DeepPINK is able to reliably choose relevant features with a controlled error rate, regardless of arbitrarily complex interactions among features. To rigorously control the false discover rate among selected features, DeepPINK relies upon the recently described model-X knockoffs framework (Barber and Candès 2015). The primary methodological novelty in DeepPINK is its deep neural network architecture, which enables application of the model-X framework.

2.3.15 *Determining feature importance*

After training, we examined each model to extract information about which features the model deemed most relevant to the given classification task. For the logistic regression models, we recorded the coefficients assigned to each feature. For XGBoost, we used the SHAP package

to calculate ``SHAP values'' for each feature at each gene (Lundberg and Lee 2017). The magnitude of the feature importance score was defined as the mean SHAP value across all genes. The sign for the feature importance score (indicating whether a feature indicates high- or low-expression) was defined by the direction of correlation between feature values and SHAP values across all genes. DeepPINK computes feature weights by multiplying the weight matrices across all layers of the deep neural network. The resulting weights can be either positive or negative, indicating the direction of correlation between features and the label. We used the squared value of the feature weight as the feature importance score.

2.4 RESULTS

2.4.1 *High- and low-expression genes display qualitative genetic and epigenetic differences*

Drawing from a variety of data sources, as described in Methods, we constructed a data set of heterogeneous gene features across each of three stages of the erythrocytic cycle (ring, trophozoite and schizont). GRO-seq measurements of nascent transcription were used to identify genes which high expression (top third) and low expression (bottom third) (Figure 2.1A).

As an initial step of data exploration, features with signal at a base-by-base level (such as ChIP-seq tracks or GC content) were visualized using aggregation plots showing the average level of signal for a feature with respect to the start codon, segregated by high-expression and low-expression genes (Figure 2.1B). These plots show expected trends, including enrichment of H3K4me3, H3K9Ac, and H2Az in highly expressed genes, as well as depletion of nucleosome occupancy in promoter regions.

Given the apparent differences in signals upstream and downstream of the start codon, we split each of these main features into two features. The ``promoter'' feature was the mean feature

signal from -500 bases up to the start codon, whereas the ``gene body" feature was the mean feature signal from the start codon to 1 kb into the coding sequence. This was done for all covalent histone modification features, H2Az composition, nucleosome occupancy, and GC content.

Similarly, for motif features, we calculated the maximum motif match log-odds score in two windows. The ``promoter" window was from -500 bases up to the start codon, while the ``gene body" region extended from the start codon up to 1000 bases into the coding sequence.

Ultimately, each ring- and schizont-stage gene was characterized by a set of 73 features, including 50 motif-based features, 14 histone modification features, 7 features characterizing local and global chromatin structure, and 2 features describing local GC content. Trophozoite stage genes used the same feature vector, but with 10 histone features removed due to missing ChIP-seq data sets in that stage. These matrices, including feature values and gene labels, are available for all three stages via the Github repository under the modelData directory. For each stage, we include two versions of the file, with and without motif features, for convenience.

2.4.2 *Machine learning models accurately distinguish between expressed and non-expressed genes*

We initially examined single features to establish a baseline of classification performance based on a simple ordering of genes by each individual feature. In this way, we generated one ROC curve for each feature (gray lines in Figure 2.2A), obtaining AUROCs as high as 0.82 for H3K4me3 gene body signal (blue line in Figure 2.2A) in classification of ring-stage genes.

Next, we compared this baseline approach against a machine learning method that integrates all of the available features. We observed, not surprisingly, that an elastic net-regularized logistic regression (“Logistic") model that integrates all features outperformed

rankings based on single features alone: the ROC curve generated by the logistic regression (red curve, Figure 2.2A) dominates all of the ROC curves generated by ranking genes using single features. We observed similar levels of performance across the three erythrocytic stages, where in a three-fold cross-validated test, the logistic regression model achieved average AUROCs of 0.868 in ring (Figure 2.2B), 0.817 in trophozoite and 0.829 in schizont.

2.4.3 *Start codons outperform TSSs for predicting transcription*

During our exploratory work using the three ``development folds" of data, we tested two different approaches for defining the start of a gene: transcription start sites (TSSs) and start codons. Surprisingly, this testing indicated that start codons are more useful than TSSs for defining the division between promoter and gene body for our predictive models. We started by using genome-wide CAGE-seq datasets to define transcription start sites for all genes (see Methods). Plots of feature scores with respect to these two types of ``start" positions ---start codons (Figure 2.1B) and TSSs (Figure 2.3A) --- qualitatively showed stronger trends between

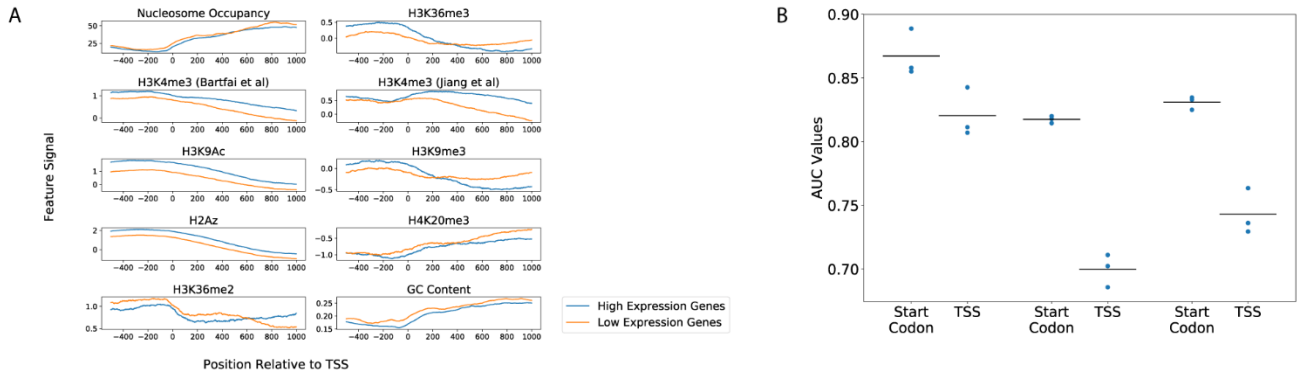

Figure 2-3: **Transcription start sites versus start codons.** A) Aggregation plots showing the average signal for features with respect to the transcription start sites of high-expression (blue) and low-expression (orange) genes. B) A plot of AUROC scores obtained for training a logistic classifier to classify genes as high- or low-expression (in the Ring, Trophozoite, and Schizont stages, left to right). The left column for each stage represents scores using promoter/gene body divided at the start codon (as was done throughout the analysis up to this point), while the right column in each stage used TSSs to divide promoter/gene body regions.

high- and low-expression genes when defining promoter/gene body splits using start codons rather than TSSs. Furthermore, when we trained classifiers to label genes using features split by either start codon or TSS, the models using promoter/gene body definitions split by start codons consistently out-performed models using TSSs (Figure 2.3B). Consequently, we focused analyses in this work on models that are split by start codons rather than TSSs.

2.4.4 *Motif features are not helpful*

A key question we aimed to address is the relative utility of the scores derived from TF motifs. Accordingly, we repeated the cross-validated testing of the logistic regression model using three different feature sets: the full set of features, a reduced set in which the TF motif PWM scores have been eliminated, and a set containing only TF motif features. This analysis suggested that the motif features did not aid in classification when combined with non-motif features, and if anything hurt the performance of our models (Figure 2.2C). Furthermore, models that used only motif features were far less accurate than models that incorporated non-motif features (Figure 2.2C). In addition, we investigated the possibility that the 500 bp window size used for promoter features may have under-utilized AP2 motifs, if relevant regulatory sequences are spread over larger upstream distances. To this end, we re-trained and evaluated new ``motif-only'' models, varying the promoter region to include either 1~kb or 2~kb of upstream sequence (instead of 500~bp, as in the original analysis). This analysis (Figure 2.2D) shows that expanding the range of the upstream window does not improve the performance of models using motif features alone, suggesting that our modest upstream window size is not missing valuable upstream regulatory AP2 family binding sites.

At this point, we considered our model development complete. Hence, all subsequent analyses incorporate two folds of data that had not been used in prior model development. Thus,

whereas previous analyses involve three-fold cross-validation on 3/5 of the data, all subsequent analyses perform two folds of a five-fold cross-validation, training on 4/5 of the data and testing on each of the two held-out test folds (see Methods for details).

2.4.5 *Different models have similar accuracies*

To determine whether our results thus far depend upon the choice of machine learning method, we also tested two additional types of models: a boosted trees ensemble (``XGBoost"), and a multilayer perceptron with two hidden layers (``DeepPINK"). Refer to “methods" for descriptions of the methods and links to further reading. These methods all demonstrated similar performance on our development folds, so we elected to carry out further analysis using all three modeling approaches in parallel. In each stage, the three models demonstrate similar AUROC performance, with a slight trend of the multi-layer perceptron model outperforming XGBoost, which in turn outperforms logistic regression (Figure 2.2E). We examined all pairwise model comparisons within each stage (DeLong test for correlated AUCs, see methods), finding three comparisons to have statistically significant differences (Figure 2.2E). However, even the differences that are statistically significant are relatively modest in absolute terms, leading us to conclude that each of these machine learning methods achieves reasonably good performance in discriminating between *Plasmodium* genes with high and low expression.

Accordingly, we used all three methods in subsequent analyses.

2.4.6 *Classification models use stage-specific features*

Having established the feasibility of predicting gene expression in *Plasmodium*, we next turned to the more interesting question: which features contribute most strongly to the performance of each classifier? By including three different types of models, we reasoned that if

multiple classification models select a similar set of informative features within a single stage, then this would suggest that those features are robust to the choice of model. Accordingly, for each model we calculated a feature importance score (see Methods) on a 0 to 1 scale, where 0 means uninformative and 1 means strongly informative. We also determined the direction of effect, indicating whether a high feature value is predictive of high or low expression. Additionally, the DeepPINK model identifies which features are informative for classification using a method that allows for explicit control of false discovery rate (FDR < 0.05, see Methods). Note that due to the exclusion of motif features from our analysis (Figure 2.2C), we do not obtain feature scores for any AP2 motifs. Given the fundamentally different methods used to assign feature importance in each of the three models, we would not expect a precise quantitative agreement in scores. For instance, elastic-net regularized regression models favor zero-valued coefficients, whereas no such sparsity-inducing behavior occurs in XGBoost or DeepPINK models.

However, we would expect that model agreement on feature importance would result in similar feature rankings. Indeed, when we calculate the Spearman correlation between all model pairings in a given stage we see a strong correlation between different models' feature orderings (Table 2.4). We note that the agreement is imperfect, particularly between XGBoost and DeepPINK models in the schizont stage (correlation = 0.533), with obvious differences between the two models in their use of H3K36me2 and H3K36me3, among others (Figure 2.4A). In general, occasional instances of disagreement between models feature attributions within a given stage (Figure 2.4A and Table 2.4) are difficult to interpret with

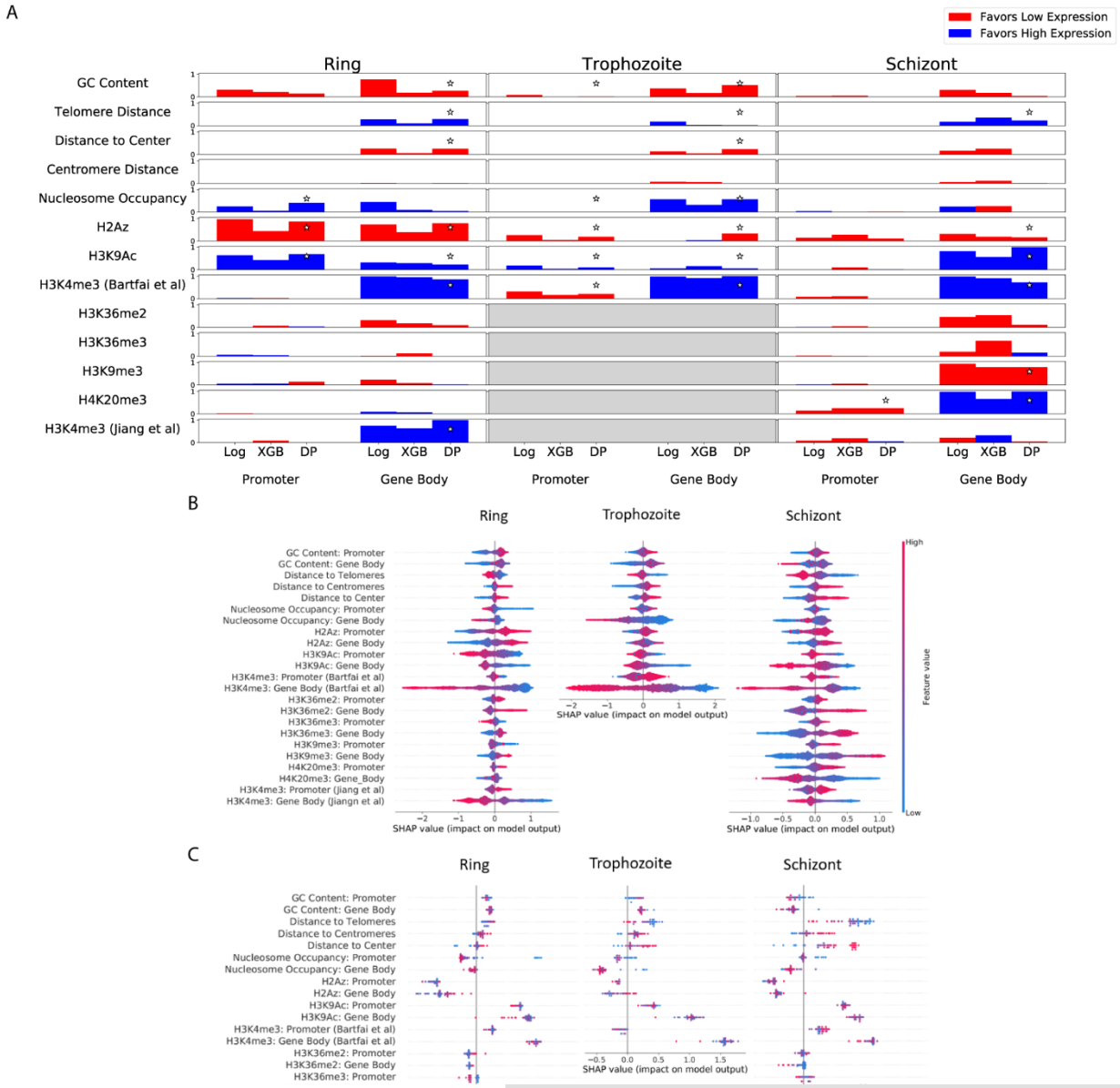

Figure 2-4: **Feature importance measures.** A) Feature importance scores assigned for different models (Log = Logistic Regression, XGB = XGBoost, DP = DeepPINK Multilayer Perceptron). See "Methods" for details. Scores were normalized to a 0-to-1 range. Bar height represents the magnitude of feature significance, while the color of bars indicates the direction of effect (Red: Higher feature value predicts high expression. Blue: Higher feature value predicts low expression.). Features using averages over "promoter" and "gene body" windows (such as ChIP-seq tracks) are split by these sub-features, while features that are not divided are not. Stars indicate features that were selected as significant using the DeepPINK model, FDR <0.05. B) SHAP values for features used in the XGBoost classifier for all genes. A positive SHAP score for a feature for a specific gene means that the value of that feature was changed that gene's classification toward "low expression," and vice-versa. C) SHAP values for features used in the XGBoost classifier for virulence genes.

confidence given the lack of a ground truth to which we can compare the models' feature attributions.

Given the observation that our models attribute similar importance to features within a single stage (Table 2.4 and Figure 2.4A), we inspected the features that the models selected as informative. All three models identify high H3K4me3 signal within the gene body as indicative of high expression in the ring stage, select high H3K4me3 signal as indicative of high expression in the trophozoite stage, and identify high H3K9Ac and H4K20me3 signal in the gene body as indicative of high expression in the schizont stage. Similarly, the three models tend to attribute consistent importance to physical chromatin features: all three models highlight the importance of gene body nucleosome occupancy in the trophozoite stage and telomere distance in the ring stage (using inferred distance based on a 3D computational model, see methods). This consistency across models and methods suggests that our approach to identifying informative features is generally robust to the differences in modeling approaches.

Model comparison	Ring features	Trophozoite features	Schizont features
Logistic vs. XGBoost	0.945	0.934	0.718
Logistic vs. DeepPINK	0.876	0.857	0.772
XGBoost vs. DeepPINK	0.821	0.775	0.533

Table 2-4: **Intra-stage consistency of model feature attributions**. Spearman correlations between all pairs of models were calculated for features within individual stages, as well as averaged across all stages.

In contrast to the high concordance among the three models, we observed low concordance among the importance of individual features across different stages of the erythrocytic cycle. For instance, H4K20me3 was highly informative for predicting a ``high expression" label in the schizont stage, but almost completely uninformative in the ring stage (data was unavailable for this mark in the trophozoite stage). To investigate the extent of this disagreement, we calculated Spearman correlations for all stage pairs for a given model type

Model	Ring vs. Trophozoite	Ring vs. Schizont	Trophozoite vs. Schizont
Logistic	.830	.632	.722
XGB	.711	.641	.365
DeepPINK	.758	.385	.565

Table 2-5: **Inter-stage consistency of model feature attributions**. Spearman correlations between all pair-wise comparisons of stages were calculated for features within individual model types.

(Table 2.5). These correlations (mean = 0.623) are notably lower than the correlations observed between different models trained within the same stage (Table 2.4, mean = 0.803). The comparatively higher consistency of feature importance within a stage versus between stages (difference = .18. p = .0176, two-sided t-test) argues that inter-stage differences are not an artifact of the model training process and suggests that distinct regulatory mechanisms may control transcription in the three different stages. However, further work and replication is required to rule out confounding issues such as batch effects between datasets for different stages. For instance, the H3K4me3 features from one source (Bártfai et al. 2010) was marked as informative for classification in the schizont stage by all three models, while H3K4me3 signal from a different source (Jiang et al. 2013) was found to be relatively uninformative, by comparison (Figure 2.4A). Such discrepancies likely stem from differences in either the data generation processes or the synchronization of parasitic stages across distinct sources.

The XGBoost model afforded an additional look at each individual features' effects on the classification of single genes. In addition to assigning feature significance and direction-of-effect at the level of the model as a whole (as in Figure 2.4A), the SHAP score for the XGBoost model can be calculated separately for each feature at each gene. Briefly, to generate classifications the XGBoost model generates a score for each gene. Suppose that correctly classified ``low-expression'' genes receive scores in the range 2-8 (on an arbitrary scale), and a particular low-expression gene has been given a score of 6, indicating that the model (correctly) predicts that the gene is at a low expression level. SHAP scores assign scores to each of the 23

features used in the prediction, such that the sum of the 23 scores adds up to 6, the final classification value. In this way, large positive SHAP scores indicate features that were important for assigning this particular gene a “high-expression" label (see Section 2.3.13 and (Lundberg and Lee 2017) for details). The resulting distribution of per-gene SHAP scores for each feature (Figure 2.4B) suggests that some features exhibit non-linear relationships between SHAP scores and expression prediction, visually observable as asymmetry in the density plots shown in Figure 2.4B. For instance, the effect of H3K36me2 gene body signal in the schizont stage model is not a simple relationship where increases in the feature value led to consistent changes in model classification (Figure 2.5A). In a histogram showing H3K36me2 distributions for high- and low-expression genes (Figure 4C), we see that over the range of -4 to -3, almost all genes are ``low-expression." This corresponds to the ``flat" region in the -4 to -3 range in the SHAP scores of Figure 2.5A, because values within this range are all treated as essentially the same by the XGBoost model. In contrast, between 0 and 1 we observe a shift in the relative abundances of high- and low-expression genes: most genes with 0 signal are high-expression, whereas most genes with 1 signal are low expression. Consequently, we see that SHAP scores have a steep slope over the 0 to 1 range, meaning that small changes in H3K36me2 have large effects on model predictions for genes within this range. Similarly, a unit change in gene body H3K4me3 intensity does not lead to a specific change in XGBoost predictions across all ranges of H3K4me3 signal (Figure 2.5B). We see a marked change in the relationship (slope) between SHAP scores and H3K4me3 signal around a score of ``0.5" (Figure 2.5B), which is at the point at which genes transition from being mostly low-expression to mostly high-expression, as seen in the distributions of Figure 2.5D.

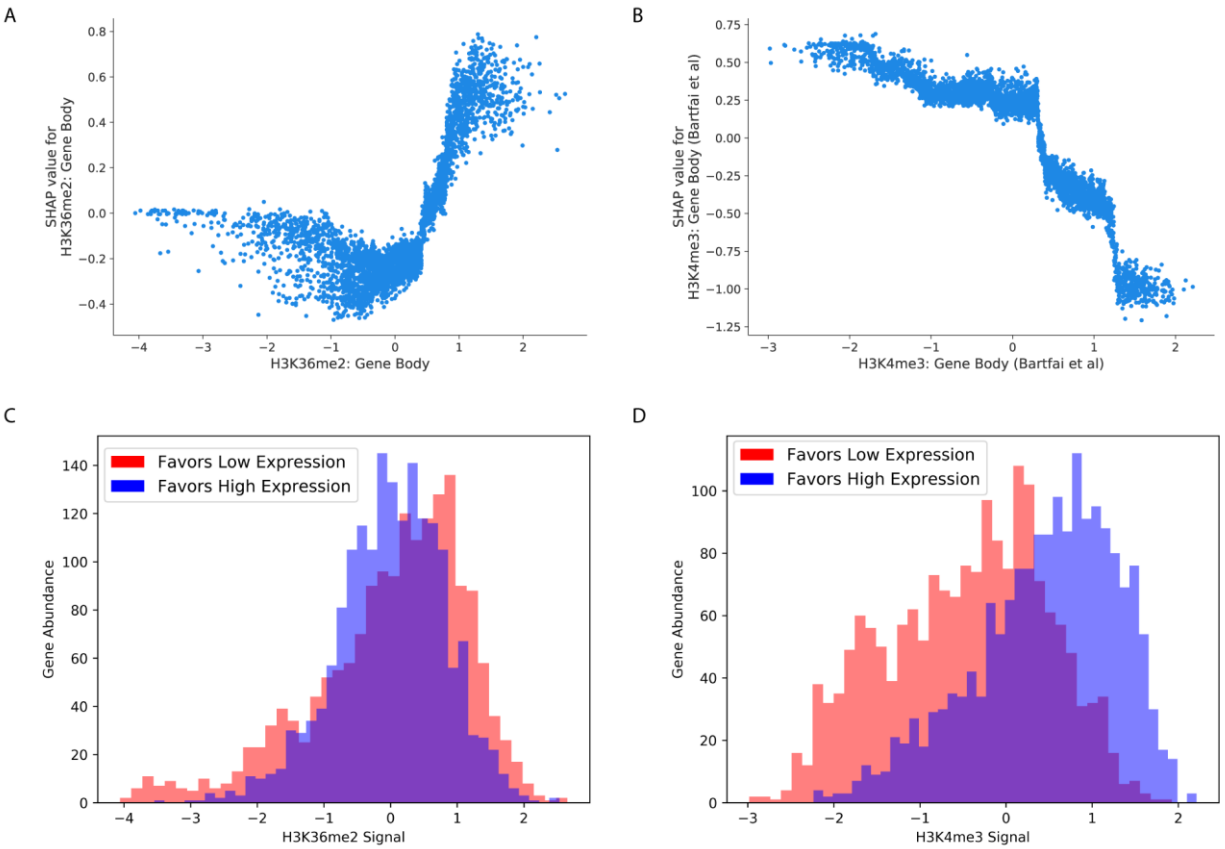

Figure 2-5: **Nonlinear relationships between feature values and SHAP values.** A) Scatter plot showing the values of H3K36me2 signal in the gene body of schizont-stage genes (x-axis) against the SHAP values assigned to those genes (y-axis). B) A scatter plot showing H3K4me3 in the gene body of schizont genes ([35], x-axis) against SHAP values for those genes (y-axis). C) Histogram showing the H3K36me2 feature value distributions for low-expression (red) and high-expression (blue) genes. D) Histogram showing H3K4me3 values for low-expression (red) and high-expression (blue) genes.

Intuitively, this observation indicates that the XGBoost model can discriminate between feature input ranges where small changes are important versus regions where small changes are insignificant. This can help capture the behavior of underlying non-linear mechanisms: for instance, high levels of H3K4me3 may indirectly help recruit pre-initiation complex components, but after a certain level H3K4me3-mediated recruitment is no longer the rate-limiting step for transcription, so further H3K4me3 deposition will not further increase polymerase activity. Non-linear models such as XGBoost and DeepPINK are able to capture

such feature-response nonlinearities, allowing for improved predictions when modeling a process with significantly non-linear mechanisms. In contrast, a linear model like logistic regression treats -4 and -3 as being exactly as different as 0 and 1, regardless of the underlying feature distribution.

We also repeated the per-gene SHAP analysis for the *Plasmodium* virulence genes, which encode a protein family that functions to anchor infected erythrocytes to the endothelium of blood vessels and are an important target for immune recognition (Flick and Chen 2004). The virulence genes are tightly regulated, with each parasite expressing exactly one of the 60 genes at a given time. In agreement with a known role for H3K9me3 in repression of virulence genes (Chookajorn et al. 2007), we find that virulence genes have large SHAP scores assigned to H3K9me3 signal. This observation demonstrates that the classification model is not only able to find genome-wide rules for classification, but also selects important features with respect to a specific subset of genes, capturing factors that are known to be important for transcriptional control of that gene family.

2.5 DISCUSSION

We developed predictive models for *Plasmodium* gene expression that yield AUC values in the range of 0.79--0.88 in cross-validated testing. These values are somewhat lower than AUC values reported from studies carried out in other eukaryotes like mouse (0.94 (Cheng et al. 2011)) or human (0.95 (Dong et al. 2012)). Many factors may contribute to this difference. For example, *Plasmodium* has a smaller number of datasets available for use as features in our models: at most six unique histone covalent modifications were used in our models, whereas 11 unique histone modification features were used in both (Cheng et al. 2011) and (Dong et al.

2012). Consistent with this, using a small feature set (five histone modifications) to classify expression in human cells resulted in a model with an average AUC of ~0.8, a value in line with the performance we observed. Furthermore, compared to human and mouse, *Plasmodium* has far fewer genes, which yields fewer examples for training our models. Additionally, the high AT-content of the *Plasmodium* genome presents a consistent challenge to generation of high-confidence genomic datasets (Ay et al. 2015), so noise in feature datasets may have led to reduced accuracy. An alternate explanation comes from the apparent abundance of genes related to post-transcriptional regulation, rather than gene-specific transcriptional control (Coulson, Hall, and Ouzounis 2004). This discrepancy has led to speculation that the most significant level of gene expression control occurs at regulation of translation, relaxing requirements for strict transcriptional regulatory programs (Coulson, Hall, and Ouzounis 2004). It is possible that a relatively low reliance on strict transcriptional control allows the parasite to tolerate high noise in transcriptional regulation, in turn leading to a system that is difficult to model accurately.

One surprising outcome was the apparently low utility of features derived from AP2 family TF binding motifs. *Plasmodium* AP2 genes are conserved proteins containing putative DNA-binding domains, homologous to the plant Apetela2/Ethylene Response Factor (AP2/ERF) DNA-binding proteins, the second largest class of transcription factors in *Arapidopsis thaliana* (Riechmann and Meyerowitz 1998). Gene expression profiling of a number of Plasmodium species as well as targeted knockout studies have demonstrated that some of these proteins are transcriptionally regulated and play key roles during developmental stages, including sexual differentiation (Painter, Campbell, and Llinás 2011). Furthermore in vitro protein binding microarray experiments prompt to the identification of DNA binding specificities for these proteins. Finally, a pronounced paucity of alternative transcription factors with DNA sequence specificity

(Campbell et al. 2010) and the presence of high-affinity AP2 motifs upstream of genes whose expression varied across the erythrocytic cycle led to the notion that these Ap2 factors could be the missing reservoir of sequence specific TFs in Plasmodium. The strikingly low value of AP2 motifs that we obtained raises three possibilities. First, this result may be indicative of the relatively low importance of local TF activity in regulating gene expression during the erythrocytic cycle. Second, we cannot completely rule out the possibility that the low value of AP2 motifs arose simply because the motifs used here are of low quality or because the way we employed the motifs (by scanning and aggregating p-values) is suboptimal. Alternately, it is possible that AP2 DNA binding requires both a TF-specific motif and a permissive epigenetic state at a given locus. DNA accessibility and epigenetic state is known to play a role in restricting TF binding in eukaryotes generally (MacQuarrie et al. 2011), with the consequence that TF motifs are an imperfect predictor of DNA binding in the absence of additional epigenomic data (Blatti et al. 2015; Duren et al. 2017). In *Plasmodium* specifically, detailed study of one TF found that the presence of a consensus motif was neither strictly necessary or sufficient for TF binding (Gissot et al. 2005). If local chromatin state affects TF binding even in the presence of a TF-specific DNA motif, the predictive value of AP2 motifs could be masked by subtle interactions with local DNA accessibility and chromatin state. Consistent with this possibility, previous models of mammalian gene expression based on sequence motifs captured a small amount of gene expression variation (Conlon et al. 2003; Bussemaker, Li, and Siggia 2001), while models using TF binding assayed by ChIP-Seq were able to predict expression with far greater accuracy (Cheng et al. 2011). This is presumably because ChIP-Seq data implicitly captures both motif presence/absence as well as epigenetic factors affecting TF binding. Clearly, an extensive collection of TF ChIP-seq data would be hugely valuable in exploring the extent to

which TFs play an active role in gene regulation in *Plasmodium* and would clarify if AP2 factors truly play a limited role in erythrocytic transcriptional regulation. Initial Chip-seq results against AP2-G2 and AP2-I, transcription factors thought to be involved in sexual development and cell invasion respectively, suggest that AP2 may interact with some promotors to either act as a repressor for AP2-G2 (Yuda et al. 2015) or activator in associates with several chromatin-associated proteins, including the Plasmodium bromodomain protein PfBDP1 for Ap2I (Santos et al. 2017).

Inspection of our trained models revealed the use of multiple types of features, from local histone modifications to high-order spatial positioning. Covalent histone modifications were consistently found to be informative features, including the designation of gene-body H3K9Ac and H3K4me3 (Bártfai et al. 2010) as statistically significant by DeepPINK FDR control in all three stages (Figure 2.4A). Furthermore, nucleosome occupancy and GC content were repeatedly identified as informative features (Figure 2.4A, Ring and Trophozoite feature use). Together, these observations indicate that nucleosome occupancy, histone modification status, and GC content all contain valuable information regarding the activity status of a locus. In addition, the gene distances to telomere cluster and nuclear center (based on 3D models from our groups' previously generated data, see methods) were also consistently informative for classification of *Plasmodium* gene expression, albeit to a lesser extent than local features such as histone modifications (Figure 2.4A). This is consistent with previous observations that *Plasmodium* expression correlates with gene spatial positioning (Ay et al. 2014), and suggests that *Plasmodium* may encode regulatory information in the 3D position of a gene, in addition to its local epigenetic state. Our findings complement previous identification of co-regulatory

relationships between functionally related genes in *Plasmodium* (Prat et al. 2011), with our analysis identifying a repertoire of epigenetic features that underpin such observed patterns.

Interestingly, in the DeepPINK model H2Az coverage in the gene body of trophozoite genes was marked as significant (FDR < 0.05) and associated with low expression. In contrast, scores assigned to this feature were close to zero for both the Logistic and XGBoost models. H2Az signal was previously reported to be almost completely absent from gene coding sequence (Bártfai et al. 2010), which makes the apparent significance of gene-body H2Az signal quite surprising. Follow-up validation would be required to see if a minimal level of H2Az truly encodes information within coding sequence, or if the identification of the feature as significant is an artifact of the DeepPINK procedure. However, previous studies in metazoan genomes have also identified H2A.Z in gene bodies. While some research groups link low levels of H2A.Z with inhibition of transcription in reconstituted nucleosomes (Mavrich et al. 2008; Thakar et al. 2010), others suggest that H2A.Z nucleosomes may facilitate transcriptional elongation (Weber, Henikoff, and Henikoff 2010). Our results support a model in which a low level of H2A.Z nucleosomes acts as a simple barrier to transcriptional elongation. However, given the general agreement between models for almost all other features (Figure 2.4A) the assignment of importance to H2Az signal by DeepPINK alone suggests that the relationship should be considered very tentative.

An inherent limitation of our analysis is that, given these data, we cannot easily separate correlations from causative relationships. This is particularly important when modeling transcription using epigenetic data, given previous evidence that some histone marks (H3K36 and H3K79 methylation) are deposited directly as an effect of Pol II elongation, rather than preceding transcriptional activation (Gates, Foulds, and O'Malley 2017). In the absence of

detailed perturbational experiments, the predictive relationships that we observe between features and expression cannot be clearly defined as directly regulatory or not.

Despite this limitation, our identification of predictive features is helpful on two fronts. First, epigenomic changes resulting from transcriptional activity can themselves serve in regulatory roles. In some species, H3K36 methylation, for instance, is deposited concurrently with transcription but serves a regulatory role thereafter, suppressing aberrant initiation of transcription within gene bodies (Gates, Foulds, and O'Malley 2017). This means that our models may identify factors important not only for regulation preceding initial activation of a locus, but also for feedback regulatory mechanisms. Second, the observed differences in selected features in distinct stages gives a clear prioritization for points in the *Plasmodium* life cycle where experimental dissection of epigenetic function would be most informative. For instance, H4K20me3 is not predictive of expression the ring stage, but is consistently associated with transcriptional repression in the schizont stage (Figure 2.4A). Whether H4K20me3 is a cause or effect of transcription, the molecular events linking this mark to transcription likely only take place---and are experimentally targetable---in the schizont stage. Our analysis specifically suggests that H3K9me3, H4K20me3, and K3K9Ac play schizont-specific regulatory roles in the erythrocytic cycle of *Plasmodium* (Figure 2.4A). This observation suggests that disruption of enzymes controlling the levels of these marks would result in schizont-specific dysregulation, either through genetic ablation or chemical inhibition. Therapeutic targeting of specific epigenetic pathways is already an active area of study in oncology (Fardi, Solali, and Farshdousti Hagh 2018) and virology (Archin et al. 2017), and future efforts applying epigenetic disruption to antimalarial regimens will benefit from our determination that the schizont stage appears to rely upon a larger variety of covalent histone modifications than other erythrocytic stages.

Analyzing XGBoost models suggested that the best solution to the classification task did not take the simple form in which a unit increase in a given feature leads to a specific, constant change in classification probability (Figure 2.4B and (Lundberg and Lee 2017)). Consistent with this, our two approaches that allow for feature interactions and non-linear feature/classification relationships, XGBoost and DeepPINK, slightly but consistently out-performed logistic regression (Figure 2.2D). However, the improvements in test AUC for the XGBoost/DeepPINK models are statistically significant in only a subset of these comparisons, and in all cases are quite modest in absolute terms. This is consistent with work in other eukaryotic genomes, where incorporating feature interactions provided minimal improvement in gene expression prediction accuracy, compared to simple linear models (Cheng et al. 2011; Dong et al. 2012). It is possible that complex models would demonstrate a more significant advantage in a regression task---such as predicting absolute mRNA abundance---rather than the binary classification task that we considered. In our case, however, it appears that models using simple additive effects, such as logistic regression, captured most of the information found within the input features.

Our work only studied factors associated with relative control of expression within a particular erythrocytic stage. This approach has the limitation of ignoring gene expression dynamics related to changes in absolute expression. In our per-stage labeling approach, 2937 genes received the same label (``high", ``low", or ``intermediate") in all three stages, 2162 were either ``high" and ``intermediate" or ``low" and ``intermediate", and the remaining 182 were labeled as ``high" and ``low" expression at least once. Figure 2.1A shows that genome-wide expression values vary widely between stages, consistent with known inter-stage variation in transcriptional activity. From this, we know that many genes are changing in absolute expression levels between stages but ending up with a “constant" expression label when

classified by relative expression. Conversely, a subset of housekeeping or otherwise constantly expressed genes may not actually vary in absolute expression themselves but end up with varying high/intermediate/low labels due to global shifts of transcription. An alternative modeling approach could build upon our analysis of intra-stage expression regulation to incorporate inter-stage expression changes and absolute expression regulation, with the aim of building a complementary picture of *Plasmodium* transcriptional control.

During model development and feature refinement, we came to the surprising discovery that placing the promoter/gene body division using start codon position was more effective than using transcription start sites (Figure 2.3B). This observation is consistent with a previous analysis in which five out of six covalent histone modifications associated with high transcription in *Plasmodium* displayed peak enrichment at the start codons of *Plasmodium* genes, while only one displayed the highest enrichment at transcription start sites (Gupta et al. 2013). Additionally, this is consistent with the observation that *Plasmodium* lacks a strongly positioned +1 nucleosome at the TSS, but that clearly positioned nucleosomes are observed at the start and end of coding sequences (Bunnik et al. 2014; Ay et al. 2015). In the future, it would be interesting to see if epigenetic information related to transcriptional control is truly encoded primarily with respect to start codons, or if technical artifacts due to the extreme AT bias in non-coding DNA upstream of start codons leads to the apparently limited information value of TSS-centered signals.

The following is adapted with minimal modifications from (Read et al. 2022)

Chapter 3. SINGLE-CELL ANALYSIS OF CHROMATIN AND EXPRESSION REVEALS AGE- AND SEX-ASSOCIATED ALTERATIONS IN THE HUMAN HEART

3.1 ABSTRACT

Sex differences and age-related changes in the human heart at the tissue, cell, and molecular level have been well-documented and many may be relevant for cardiovascular disease. However, how molecular programs within individual cell types vary across individuals by age and sex remains poorly characterized. To better understand this variation, we performed single-nucleus combinatorial indexing (sci) ATAC- and RNA-Seq in human heart samples from nine donors. We identify hundreds of differentially expressed genes by age and sex. Sex dependent alterations include pathways such as TGFβ signaling and metabolic shifts by sex, evident in both transcriptional alterations and differing presence of transcription factor (TF) motifs in accessible chromatin. Age was associated with changes such as immune activation-related transcriptional and chromatin accessibility differences, as well as changes in the relative proportion of cardiomyocytes, neurons, and perivascular cells. In addition, we compare our adult-derived ATAC-Seq profiles to analogous fetal cell types to identify putative developmental-stage-specific regulatory factors. Finally, we train predictive models of cell-type-specific RNA expression levels utilizing ATAC-Seq profiles to link distal regulatory sequences to promoters, quantifying the predictive value of a simple TF-to-expression regulatory grammar and identifying cell-type-specific TFs.

3.2 INTRODUCTION

Profound alterations in cardiac function and disease risk have long been evident at the level of individuals' traits such as sex (Beale et al. 2018) and age (Steenman and Lande 2017). For example, female hearts exhibit more modest declines in cardiomyocyte numbers over time than males (Olivetti et al. 1995) and display distinct vascular elasticity properties (Redfield et al. 2005), while aged hearts display ventricular hypertrophy, tissue stiffening, and inflammation (Steenman and Lande 2017; Ferrucci and Fabbri 2018). However, there is substantial uncertainty in the exact molecular and cellular hallmarks - much less causal mechanisms - of these clinically evident, consequential differences. A robust understanding of those molecular processes could set the stage for personalized therapeutic intervention.

To achieve cell-type-resolved but high-throughput measurements of cardiac biology, single cell methods have been employed in numerous studies of human hearts and model organisms. In humans, these analyses profiled the diversity of cardiac cell types and subtypes (Litviňuková et al. 2020; Tucker et al. 2020) and generated genome-wide maps of cell-type specific regulatory programs (Hocker et al. 2021). In model organisms, single cell approaches have not only generated atlases of healthy tissue (Han et al. 2018; Tabula Muris Consortium et al. 2018) but have also been used in controlled experiments to dissect the alterations occurring in processes such as aging (Tabula Muris Consortium 2020) and heart disease (Ruiz-Villalba et al. 2020; Farbehi et al. 2019; Y. Zhang et al. 2019). Similar approaches have begun to profile clinically important contrasts in human samples directly, such as identifying a handful of transcripts that vary by age in the healthy human heart (Tucker et al. 2020), variation in myeloid cell abundance in age (Koenig et al. 2022), and alterations during disease in single-cell ATAC-Seq (Hocker et al. 2021) and RNA-Seq (Koenig et al. 2022) data. Further analyses and larger datasets of the human heart may unlock more

extensive insights into how alterations in transcriptional and epigenetic states characterize variation between individuals and advance our understanding of the genomic programs regulating cells.

Chromatin regulation represents a significant element of specialized cell function within or between conditions. During development, transcription factors play variable roles over the course of cardiac development (Akerberg et al. 2019; K. Zhang et al. 2021). Of clinical concern, individual transcription factors may play decisive roles in diseases such as cardiac fibrosis (Alexanian et al. 2021) while genetic variation may act through regulatory mechanisms to affect disease risk and individual variation (Y. Wang and Wang 2019). Parallel advances in quantitative models of gene expression (Avsec et al. 2021; Agarwal and Shendure 2020; J. Zhou et al. 2018) and extensive generation of epigenetic datasets in primary human hearts (Hocker et al. 2021; K. Zhang et al. 2021; Domcke et al. 2020) bode well for the utility of further, diverse epigenetic datasets in revealing intra- and inter-state regulatory programs.

To extend knowledge of molecular cell-type-specific cardiac processes between and within individuals, we generated and analyzed matched single-cell ATAC-Seq (117,738 cells) and RNA-Seq (89,404 cells) datasets from 15 samples spanning 9 individuals. As a resource, our dataset contributes substantially to the catalog of single-cell profiles of the human heart. The number of individuals profiled combined with a hierarchical mixed model regression approach allows us to resolve age- and sex-dependent transcriptional and chromatin accessibility changes apart from confounding by donor-level variation. We find that transcriptional and regulatory programs display widespread variation by these covariates, observing both cell-type-specific and largely pan-cell type alterations. For example, sex was associated with alterations in transcriptional signatures of oxidative phosphorylation as well as differing accessibility at ATAC-Seq peaks

containing motifs of TFs known to regulate metabolic rewiring. Furthermore, we employ ATAC-Seq data to identify putative life-stage specific TFs, finding indications of adult-specific activity by RFX family TFs in adult vascular endothelium and macrophages. Finally, we develop cell-type-specific gene expression models that utilize informative distal regulatory sites to account for approximately a quarter of transcriptional variation using a simple TF motif regulatory code.

3.3 METHODS

3.3.1 *Tissue Collection*

This study complies with all relevant ethical regulations and was approved by the University of Washington Institutional Review Board (STUDY00002144). Informed consent was obtained prior to collection of human tissues. No compensation was provided for participation. Collected samples were absent of evidence of disease upon review by study clinicians. Details regarding the collection are available on protocols.io (Lin and Lin 2020, 1970).

3.3.2 *Single-nucleus Library Generation*

Nuclei for sci RNA-Seq were extracted from frozen, powdered heart tissue. 200-250 mg of frozen tissue was powdered while frozen, then dissociated using a Gentle MACS Tissue Dissociator at 4C using 5 mL of ice-cold lysis/fixation buffer containing 10 mM sodium phosphate (pH 7.2), 3 mM MgCl2, 10 mM NaCl, .02% Triton X-100, 5% glutaraldehyde, 1% DEPC, 10 mM ribonucleoside vanadyl complex (NEB). Dissociated tissue was filtered through a 70 uM cell strainer on ice and washed with an additional 5 mL of ice-cold lysis/fixation buffer. The buffer/nuclei mixture was then incubated for 15 minutes at 4C in a rotating 15 mL Falcon Tube. Nuclei were pelleted by centrifugation at 600 RCF for 8 minutes at 4C. Supernatant was decanted, then nuclei were resuspended in 1 mL of nuclei suspension buffer (NSB) containing 10

mM Tris HCl, pH 7.4. 10 mM NaCl, 3 mM MgCl2, 1% SuperaseIn, 1% bovine serum albumin (BSA) solution (NEB, 20 mg/mL). Nuclei were pelleted at 600 RCF for 5 minutes at 4C, and supernatant was decanted. Nuclei were resuspended in 100 uL of NSB per aliquot, then snap-frozen with liquid nitrogen.

Libraries were generated using a 3-level sci RNA-Seq workflow (Cao et al. 2019). The workflow was modified to add a FACS sorting step following ligation to minimize background RNA levels, with a detailed workflow available at protocols.io (https://www.protocols.io/view/3-level-sci-rna-seq-with-facs-dm6gpw255lzp/v1). Libraries were sequenced using an Illumina Nextseq 500 high output sequencing kit.

Nuclei for sci ATAC-Seq were extracted from powdered, frozen tissue and fixed as in previous work (Domcke et al. 2020). Libraries were generated using a 3-level sci ATAC-Seq workflow (Domcke et al. 2020) and sequenced using an Illumina Novaseq.

3.3.3 *Single-nucleus RNA-Seq Analysis*

Raw sequencing output was processed using a pair of Nextflow processing pipelines available at https://github.com/bbi-lab/bbi-dmux (handling sample demultiplexing) and https://github.com/bbi-lab/bbi-sci (handling assignments of reads to cells, filtering, alignment, and cell-by-gene matrix generation).

Analysis of single-cell RNA-Seq data was performed using Monocle 3 (Cao et al. 2019). Cells were filtered by discarding any with unique molecular identifiers (UMIs) less than 100, mitochondrial RNA percentage greater than 10, or a Scrublet doublet score (Wolock, Lopez, and Klein 2019) of over .2. A 2-dimensional UMAP representation (McInnes, Healy, and Melville 2018) of cells was found after using mutual nearest neighbors alignment (Haghverdi et al. 2018)

to align by sample. Cell type assignments were made manually based on expression of marker genes in UMAP clusters (Figure 3.1D).

3.3.4 *Single-nucleus ATAC-Seq Analysis*

Sequencing output was processed using a pair of Nextflow processing pipelines available at https://github.com/bbi-lab/bbi-sciatac-demux (handling demultiplexing) and https://github.com/bbi-lab/bbi-sciatac-analyze (assigning reads to cells, aligning reads, calculating peaks, finding motif occurrences in peaks, and generating cell x peak matrices). Analysis of single-cell ATAC-Seq data was performed using Monocle 3 (Cao et al. 2019). Cells were filtered by discarding any with unique molecular identifiers (UMIs) less than 1000, fractions of reads in TSS (FRIT) of less than .08, fractions of reads in peaks (FRIP) less than .2, or a doublet likelihood of greater than .5 90. Gene activity scores were calculated using ArchR (Granja et al. 2021) using default settings. Cell-by-gene activity score matrices were then used to generate a Monocle CDS object. The ATAC-Seq data was then aligned with the filtered RNA-Seq data using Harmony (Korsunsky et al. 2019) based on all genes shared between RNA and ATAC CDS objects, and a new UMAP embedding was generated based on the corrected PCA coordinates of both datasets after Harmony correction. Based on UMAP coordinates in this new embedding, ATAC-Seq cells were labeled using a k-nearest neighbor transfer from the k=7 nearest RNA-Seq cells (using cell assignments described above for RNA-Seq data).

3.3.5 *Differential motif abundance testing in accessible peaks*

The presence of TF motifs in peaks was calculated based on the presence of any motif occurrence in the peak DNA sequence below a p-value cutoff of 1e-7 using MOODS (Korhonen et al. 2009). Motif count x cell matrices were then made by multiplying a motif (rows) x peaks

(columns) matrix with a peak (rows) x cell (columns) matrix, generating a motif-count x cell matrix where each entry corresponded to the number of peaks accessible in a given cell that contained a given TF motif.

To test for motif abundances that varied by a function of donor covariates (age and sex), testing was run separately for all cells of a single cell type. Testing for motif counts was done using a GLMM fit using the lme4 package ("Linear Mixed-Effects Models Using 'Eigen' and S4 [R Package Lme4 Version 1.1-28]" 2022), using a negative binomial model with sample donor as a random effect, as well as fixed effects of anatomical site, donor age, and sex. Multiple testing correction was performed with the Benjamini-Hochberg procedure (Benjamini and Hochberg 1995). This modeling approach is available in current releases of Monocle 3 (Cao et al. 2019).

For testing of cell type-specific motif enrichments (Fig. 1F), testing was run for all cells at once. To test for motifs enriched in a specific cell type, all cells were assigned a dummy variable valued as '1' for cells that are from the type being tested, and '0' for all others. Testing was then run using a GLMM fit using the lme4 package 96, using a negative binomial model with sample donor as a random effect, as well as fixed effects of the cell-type-dummy variable, anatomical site, donor age, and sex. Multiple testing correction was performed with the Benjamini-Hochberg procedure (Benjamini and Hochberg 1995).

3.3.6 *Differential expression testing*

DE testing used a GLMM fit using the lme4 package 96, using a negative binomial model with sample donor as a random effect, as well as fixed effects of anatomical site, donor age, and donor sex. This modeling approach is available in current releases of Monocle 3. Multiple testing

correction was performed with the Benjamini-Hochberg procedure (Benjamini and Hochberg 1995).

Gene set enrichment analysis tested for enrichments by age or sex within 50 Hallmark Pathways accessed from the MSigDB collection (Liberzon et al. 2015) accessed through the msigdbr R package. Testing used the fgsea package (Sergushichev 2016) and multiple testing correction was performed with the Benjamini-Hochberg procedure (Benjamini and Hochberg 1995).

3.3.7 *Adult versus fetal enrichment comparisons*

Enrichments for TF motifs in accessible chromatin of fetal cell types was accessed at https://descartes.brotmanbaty.org/bbi/human-chromatin-during-development/ (see "Motif enrichment across cell types" section for download link). Enrichments in adult cell types were calculated as described above under "Differential motif abundance testing in accessible peaks". Comparisons were made between the following adult-to-fetal matchings: "Cardiomyocyte" and "Cardiomyocytes"; "Vascular Endothelium" and "Vascular endothelial cells"; "Endocardium" and "Endocardial cells"; "Macrophage" and "Myeloid cells"; "Perivascular Cells" and "Smooth muscle cells"; "Fibroblasts" and "Stromal cells"; "Adipocytes" and "Epicardial fat cells"; "Neuronal" and "Purkinje neurons"; "T Cells" and "Thymocytes".

For each comparison, plots (Figure 3.4) were calculated using only motifs that were significantly enriched in either adult or fetal data at an FDR cutoff of .1. Outliers were selected based on qualitative divergence from broad cell type correlations in enrichments between fetal and adult cells.

3.3.8 *RNA Expression Predictive Modeling*

First, pseudo-bulk expression levels were calculated by pooling all UMIs for all genes for cells within a particular cell type. These were used to quantify the transcripts per million for each gene. Log2(TPM) was then used as the RNA expression level to be predicted for a particular gene/cell type pair.

To link distal sites to promoters, we ran Cicero (Pliner et al. 2018) to quantify covariance among peaks across all cell types. To link distal sites to genes, we first defined any peaks that intersected a defined window around the TSS (this region size was a hyperparameter set through performance on a validation set, see below). Then, any peaks outside the promoter set of peaks that were linked with a co-accessibility score greater than some cutoff (a hyperparameter). Motifs from the JASPAR database (Sandelin et al. 2004) "2018 Non-redundant Vertebrates" motif set were determined using FIMO (Grant, Bailey, and Noble 2011) at varying p-value cutoffs. For models using promoter sequence only, features would be a binary value for if one or more motif occurrences was found in the promoter sequence below a p value cutoff. For models using promoter and distal sequence, features were binary values for if a motif occurred in the promoter or distal regions.

RNA expression for protein-coding genes was predicted with an elastic net linear model using motif presence/absence as features. Data was divided into train, test, and validation sets at the level of chromosomes (approximately an 80/10/10 split in gene numbers) for hyperparameter setting. First, promoter size and motif p values were varied (Promoter sizes upstream/downstream of TSS were 1000/200, 1500/500, and 5000/2000. P values tested were 1e-4, 1e-5, 1e-6). Models were trained on the training set setting l1/l2 penalties by internal cross-validation, then evaluated on the validation set. The best average performance occurred using a p

value cutoff of 1e-4 with a promoter of 1500/500 bases upstream/downstream. Holding that promoter region size constant, we trained models varying the cicero co-accessibility cutoff to link a distal site, maximum number of distal sites to link, and window size of DNA bases to scan centered at a linked peak. We tested combinations of Co-accessibility cutoffs of .015, .035, .05; max distal sites of 5, 10, or 20; distal sit size of 600 or 1,000 bases, motif p value cutoffs of 1e-4, 1e-5, and 1e-6. Models were trained on the training set setting l1/l2 penalties by internal cross-validation, then evaluated on the validation set (Figure 3.5B). Optimal performance was obtained using a co-accessibility cutoff of .015, a maximum of 5 distal sites, 1000 bp distal site windows, and a motif cutoff of 1e-4. Those parameters were then set for use in training a model for evaluation on the test set. L1/l2 penalties were set by internal cross-validation on a pooled training + validation set, then a model was trained using those penalties and the best hyperparameters found earlier. The model was then evaluated on the test set (Figure 3.5C). Finally, holding those hyperparameters constant, a final model was trained using all three train, validation, and test sets. The coefficients of this final fit model are reported in Figure 3.5G.

3.3.9 *Data Availability*

Data is available through the HuBMAP consortium website's data portal (https://portal.hubmapconsortium.org/).

3.4 RESULTS

3.4.1 *Single cell ATAC- and RNA-Seq library generation and cell annotation*

We collected heart samples from nine healthy adult donors (Table 3.1) with hearts collected on hemodynamic support to eliminate warm ischemic time (see "Methods" section). Samples

represented four anatomical sites from the heart, though most were collected from the heart apex or left ventricular wall. In total, we prepared 15 samples for single-cell analysis, with each sample representing a single anatomical site in a particular donor. We powdered frozen tissue and split portions into aliquots for appropriate nuclei isolation and fixation for ATAC-Seq and RNA-Seq separately (Figure 3.1A).

Sample Name	Donor ID	Donor Age	Donor Sex	Anatomical Site	ATAC	RNA
W134.Apex	W134	43	F	Apex	Yes	Yes
W135.Left.Ventricle	W135	60	M	Left.Ventricle	Yes	Yes
W136.Apex	W136	43	M	Apex	Yes	Yes
W136.Left.Ventricle	W136	43	M	Left.Ventricle	Yes	Yes
W137.Apex	W137	49	F	Apex	Yes	No
W139.Left.Ventricle	W139	45	M	Left.Ventricle	Yes	Yes
W139.Right.Ventricle	W139	45	M	Right.Ventricle	Yes	Yes
W139.Septum	W139	45	M	Septum	Yes	Yes
W139.Apex	W139	45	M	Apex	Yes	Yes
W142.Apex	W142	55	F	Apex	Yes	Yes
W144.Apex	W144	53	M	Apex	Yes	Yes
W144.Left.Ventricle	W144	53	M	Left.Ventricle	Yes	Yes
W145.Apex	W145	51	M	Apex	Yes	Yes
W145.Left.Ventricle	W145	51	M	Left.Ventricle	Yes	Yes
W146.Left.Ventricle	W146	25	F	Left.Ventricle	Yes	Yes

Table 3-1: **Donor metadata.**

We generated single-nuclei RNA-Seq libraries using 3-level sci RNA-Seq (Cao et al. 2019). We modified the nuclei isolation protocol to use additional RNase inhibitors, mechanical dissociation of tissue, and 5% glutaraldehyde for tissue fixation (see "Methods" section). Additionally, in order to reduce background RNA levels that commonly contribute noise to single-cell RNA-Seq data (Young and Behjati 2020) we included a FACS sorting step following ligation. One sample failed completely (donor "W137") but the remaining 14 samples yielded 89,404 nuclei after doublet removal and filtering. A UMAP embedding of all transcriptomes contained numerous clearly separated clusters. Examination of marker genes revealed that clusters corresponded to specialized cells of the human heart including cardiomyocytes, fibroblasts, macrophages, and

endothelial cells (Figure 3.1B, C). The RNA data also clearly resolves rare cell populations such as adipocytes, neuronal cells, mast cells, and B cells (Figure 3.1B).

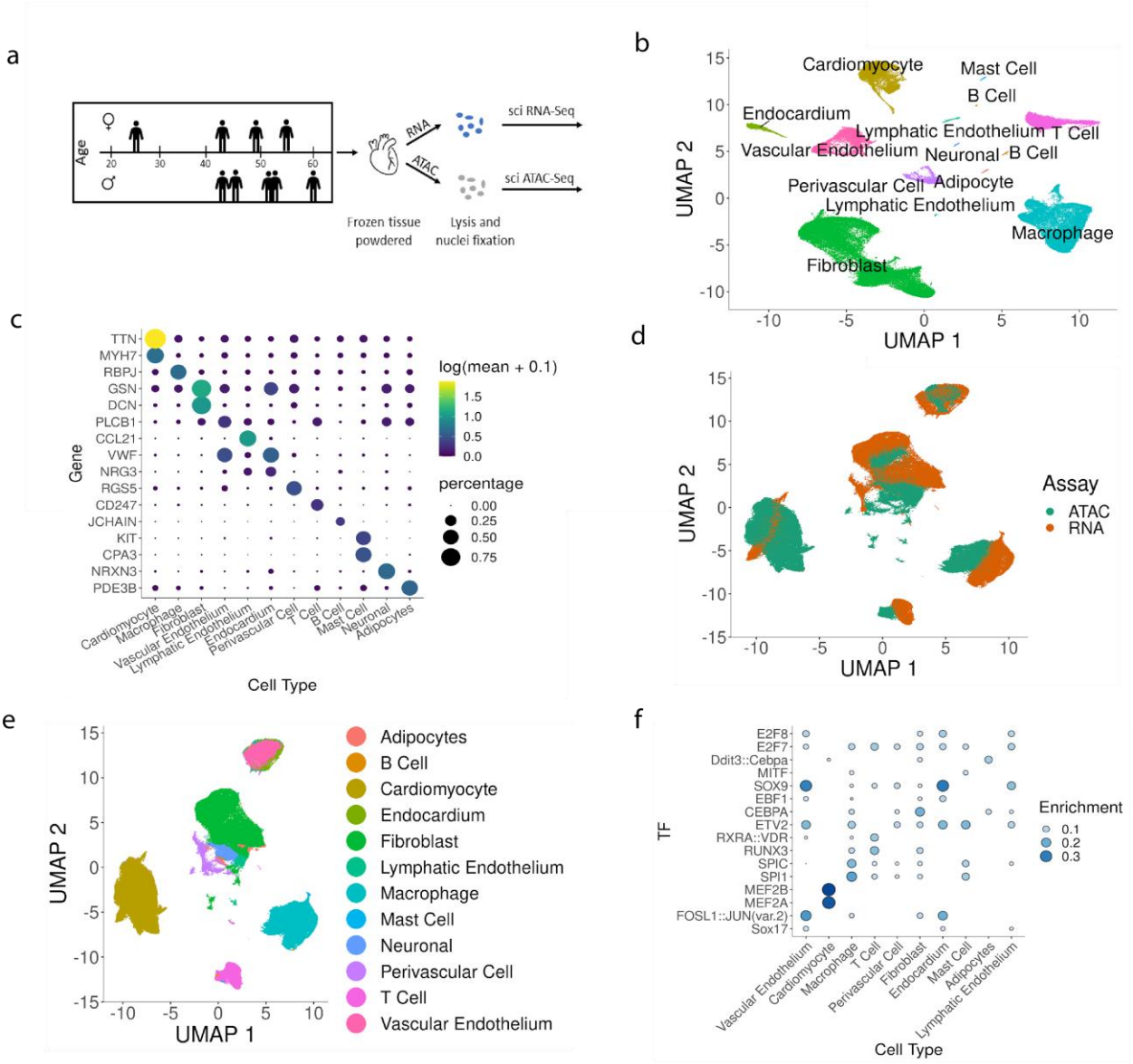

Figure 3-1: Overview of datasets. A) Experimental setup. B) UMAP of RNA-Seq data, coloring by cell type. C) Plot of marker gene expression by cell type, showing both mean expression (dot color) and the proportion of each cell type with nonzero expression for that marker (dot size). D) UMAP of a co-embedding of snRNA- and ATAC-Seq data, coloring by data type. E) UMAP of co-embedded RNA and ATAC data, coloring by the assigned cell type F) Enrichment of TF motifs in the accessible peaks of ATAC-Seq cells

Separately, we prepared single-nuclei ATAC-Seq data from powdered frozen tissue using 3-level sci ATAC-Seq (Domcke et al. 2020), generating 117,738 ATAC-Seq profiles after filtering and doublet removal. Single-cell ATAC seq data is more difficult to annotate than RNA data because open chromatin around a gene doesn't always indicate that gene is robustly expressed (Cusanovich et al. 2018). Due to this difficulty in defining cell types, we used a co-embedding approach to find a low-dimensional embedding of RNA- and ATAC-Seq data simultaneously (Figure 3.1D), then transferred cell-type labels from RNA to ATAC profiles using a k-nearest neighbor classifier (Figure 3.1E; see "Methods" section). Using these assignments, within the ATAC-Seq data we identify strong enrichments for expected cell-type-specific transcription factors (Figure 3.1F) such as MEF2 family transcription factors in cardiomyocytes, SPI1 (also known as PU.1) in macrophages, and CEBPA in fibroblasts, in agreement with recent analyses of adult single-cell ATAC-seq data in adult human hearts (Hocker et al. 2021).

3.4.2 *Sex corresponds to cell-type-specific differences including TGF-β signaling and metabolic alterations*

We explored our single cell RNA-Seq data to identify transcriptional changes associated with age or sex within individual cell types. Commonly-used single-cell differential expression methods using fixed effect regression models do not properly account for inter-sample variation and their mis-application to datasets such as ours - in which cells are not statistically independent when derived from the same donor - can dramatically inflate false discoveries (Zimmerman, Espeland, and Langefeld 2021; Squair et al. 2021). Consequently, we used a mixed effect modeling framework to test for differential expression (see "Methods" section). This approach allowed us to test for variation by sex and age while controlling for expression differences due to anatomical site (as a fixed effect) and donor (as a random effect). We find dozens to hundreds of differentially

expressed genes by age and sex, depending on the cell type analyzed (Figure 3.2A). Most DE genes are found in relatively abundant cell types, and we note particularly large numbers of changes by age in cardiomyocytes as well as many differences by sex in fibroblasts, macrophages, and vascular endothelium. For example, in vascular endothelial cells (Figure 3.2B) we observe large differences between men and women in expression of *PROM1* (also known as *CD133*), a marker for endothelial progenitor cells (Rufaihah et al. 2010) and proliferative vascular endothelium (Sekine et al. 2016). As transplanted *PROM1+* cells promote vascular regeneration in a mouse model of ischemic heart injury (Rufaihah et al. 2010), altered *PROM1* levels by sex may indicate variation in proliferative endothelial populations. DE genes by sex in vascular endothelium also include *IL1R*, a receptor for pro-inflammatory IL1 signaling that is a candidate therapeutic target pathway in cardiovascular diseases including acute myocardial infarction (Abbate et al. 2020); *ALPL*, an alkaline phosphatase that promotes cardiac fibrosis (Gao et al. 2020); and *ACVRL1* (*ALK1*), a TGFβ superfamily co-receptor that causes vascular malformations and hemorrhage upon depletion (Tual-Chalot et al. 2014).

To summarize high-level changes within sex- and age-specific variation, we tested for enrichment in up- and down-regulated gene sets. Statistically significant differences between male and female hearts were evident across several cell types (Figure 3.2C). One recurrent alteration was a decrease in TGFB hallmarks across several cell types. We observe decreased expression of target genes of TGFB signaling in male fibroblasts and macrophages as well as decreased hallmarks of epithelial-to-mesenchymal transition - a common downstream consequence of TGFB activity (J. Xu, Lamouille, and Derynck 2009) - in macrophages, vascular endothelial, and perivascular cells (Figure 3.2C). Additionally, we find statistically significant changes in genes important in various aspects of cell metabolism, including decreased expression of cholesterol

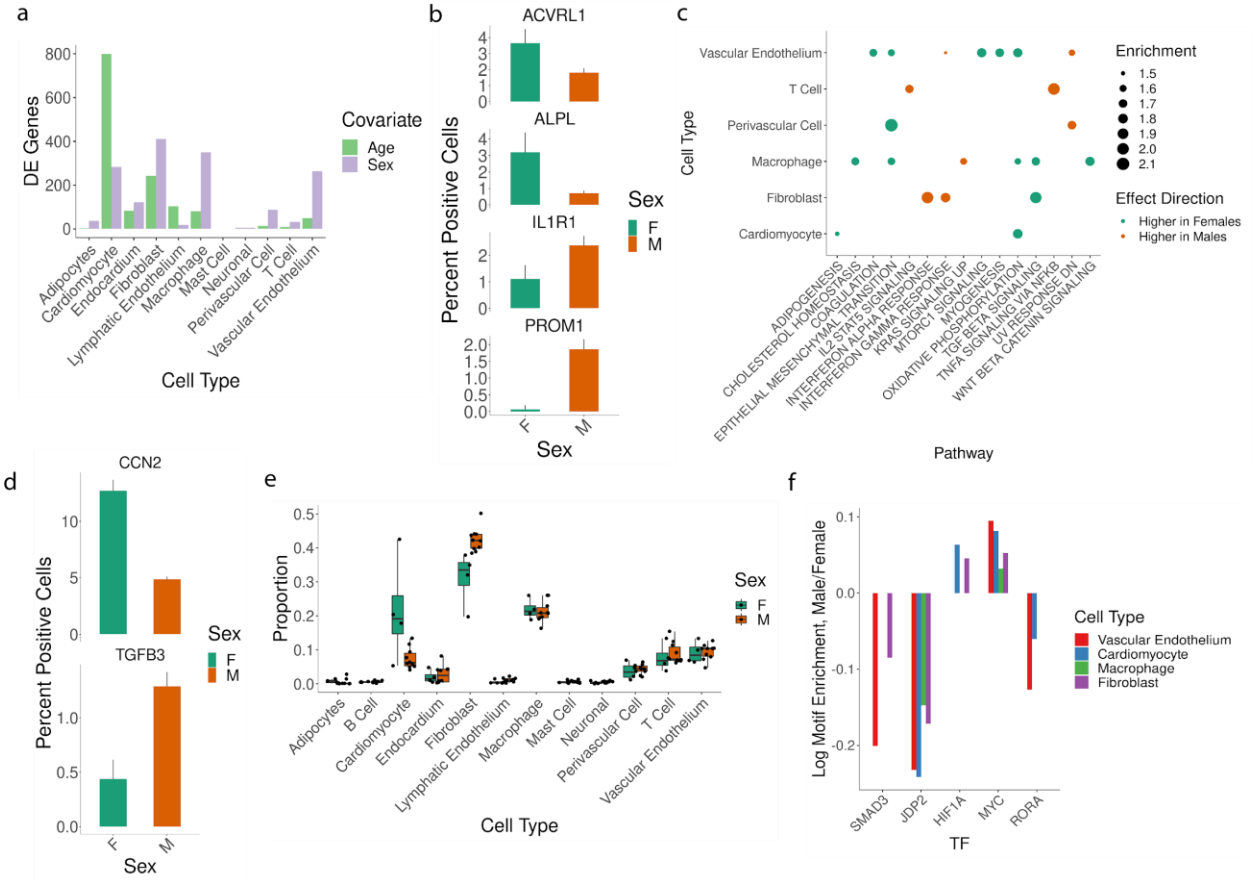

Figure 3-2: Alterations by sex in the heart. A) Total differentially expressed genes by age or sex (FDR = .1). B) Vascular endothelial cells positive for expression of four genes that were DE by sex. C) Enrichment of biological processes by sex. Pathways "Enriched in Females" show higher-than-expected expression in female cells after controlling for other covariates, and vice-versa. D) Fibroblasts positive for expression of two genes that were DE by sex. E) The proportion of total cells per sample that were classified as each cell type. F) Enrichment for counts of TF motifs in the accessible peaks of cell types as a function of sex. Bars shown correspond to statistically significant coefficients (FDR = .1).

metabolism-associated transcripts in macrophages and decreased oxidative phosphorylation-related transcripts in male cardiomyocytes, vascular endothelium, and macrophages (Figure 3.2C). We also observe cell-type specific alterations, such as an increase in hallmarks of IL2 and TNFA

signaling in male T cells (Figure 3.2C) consistent with elevated soluble inflammatory signaling (Mehta, Gracias, and Croft 2018; Ross and Cantrell 2018).

We next examined whether our data contained further clues to the source of these pathway-level changes. We find statistically significant alterations in expression levels of *TGFB3* in fibroblasts (Figure 3.2D), but counter-intuitively expression levels are higher in men than women. Apart from altered *TBFB3* in fibroblasts we do not detect significant decreases in RNA levels of *TGFB1*, *TGFB2*, or *TGFB3* in any other cell type, but potentially find other between-sex differences that could affect TGFB signaling. For example, in male fibroblasts we do find decreased levels of *CCN2*, a promoter of TGFB signaling (Nakerakanti, Bujor, and Trojanowska 2011).

As it is also plausible that differences in the abundance of cell types with active roles in intercell communication could underpin male/female differences, we examined the cell type proportions of samples by donor sex. Broadly, cell type proportions were consistent between the sexes. However, we observe some differences in cell proportions by sex including as a statistically significant increase in fibroblast proportions in male donors (Figure 3.2E) using a beta-binomial model accounting for variation due to sex, age, and anatomical site.

In order to additionally study regulatory programs whose activity differed by donor sex, we looked for alterations in TF motifs accessibility between male and female cells in order to validate and expand upon the observed transcriptional differences. Consistent with our observation of decreased TGFB signaling in transcriptional data, we see reduced abundance of motifs for SMAD3 - a downstream effector of canonical TGFβ signaling - in male fibroblasts, vascular endothelial cells, and macrophages (Figure 3.2F). We further see statistically significant decreases in accessible motifs corresponding to JDP2, a transcriptional repressor tied to alteration of TGFB1

induced EMT and fibrosis (Heger et al. 2018; Tsai et al. 2016). Consistent with decreased expression of hallmarks of oxidative phosphorylation in males (Figure 3.2F), we detect sex-specific changes across multiple cell types in HIF1A, MYC, and RORA, TFs known to promote glycolysis over oxidative phosphorylation (Rodríguez-Enríquez et al. 2019). These ATAC-Seq-based analyses align with the transcriptional decreases in TGFβ signaling and oxidative phosphorylation genes we observed in male cells and identify putative regulatory mediators of sex-specific distinctions.

3.4.3 *Immune activation increases with age through multiple pathways*

We next explored cell-type specific changes in the expression of hundreds of genes that varied by donor age (Figure 3.2A). To understand the broad changes these alterations represent, we again tested for enrichments in gene sets for age-related expression changes within individual cell types. We found a variety of alterations, including changes in several metabolic and cell-signaling pathways (Figure 3.3A). We again observed differences in TGFB signaling genes, with increased TGFB hallmarks increasing in aged fibroblasts and epithelial-to-mesenchymal transition-associated transcripts elevated in aged fibroblasts, macrophages, endocardial cells, and cardiomyocytes (Figure 3.3A**)**. Age is associated with an increase in several immune pathways across several cell types, including hallmarks of inflammation in fibroblasts and macrophages and increased interferon response in macrophages and vascular endothelium (Figure 3.3A). Statistically significant alterations include age-dependent increases in fibroblast expression of chemerin receptor *CMKLR1* (Figure 3.3B), a receptor that promotes inflammatory responses (Ho et al. 2010), mediates macrophage retention (Hart and Greaves 2010), and is positively correlated with atherosclerotic lesions (Kostopoulos et al. 2014).

To see if age-dependent alterations in immune activation hallmarks were evident at the level of chromatin remodeling, we tested for motif enrichment in accessible peaks as a function of age. We detected enrichment of IRF1 and IRF7 motifs in accessible peaks of cardiomyocytes, fibroblasts, macrophages, and vascular endothelial cells (Figure 3.3C) consistent with increases in interferon response pathways observed in our transcriptional data (Figure 3.3A). In addition, we observe a statistically significant decrease in accessible motifs for NFKB2 (Figure 3.3D), a central mediator of inflammatory signaling (Oeckinghaus and Ghosh 2009). These changes in accessibility of motifs corresponding to key mediators of immune activation corroborate observed transcriptional changes in immune-related pathways (Figure 3.3A).

Given heightened immune activation with age (Figure 3.3A,B), we wondered if increased cell senescence - a state of arrested proliferation with release of inflammatory mediators (Coppé et al. 2010) - was evident in our samples. Senescence occurs across organs (M. J. Zhang et al. 2021) and is suspected to play a role in age-related susceptibility to a host of cardiovascular diseases (Shimizu and Minamino 2019). While a clear understanding of senescence markers and causative mechanisms remains elusive (Shimizu and Minamino 2019), individual markers have been implicated in cardiac health. For example, the senescence marker p53 (Shimizu and Minamino 2019) promotes inflammation in a mouse pressure overload model (Gogiraju et al. 2015) while p53 knockout reduces age-related cardiac dysfunction in mice (Mak et al. 2017). In all cell types we do not detect statistically significant transcriptional changes of pro-senescence regulators *P53* or *CDKN1A* (p16) (Coppé et al. 2010). However, we find increased abundance with age in accessible peaks of cardiomyocytes and macrophages for CUX1 motifs (Figure 3.3C), a TF linked to senescence in vasculature by promoting transcription of p16 (G. Li 2021), a cyclin dependent kinase that promotes cell cycle arrest in senescence (X. Zhang et al. 2012). We

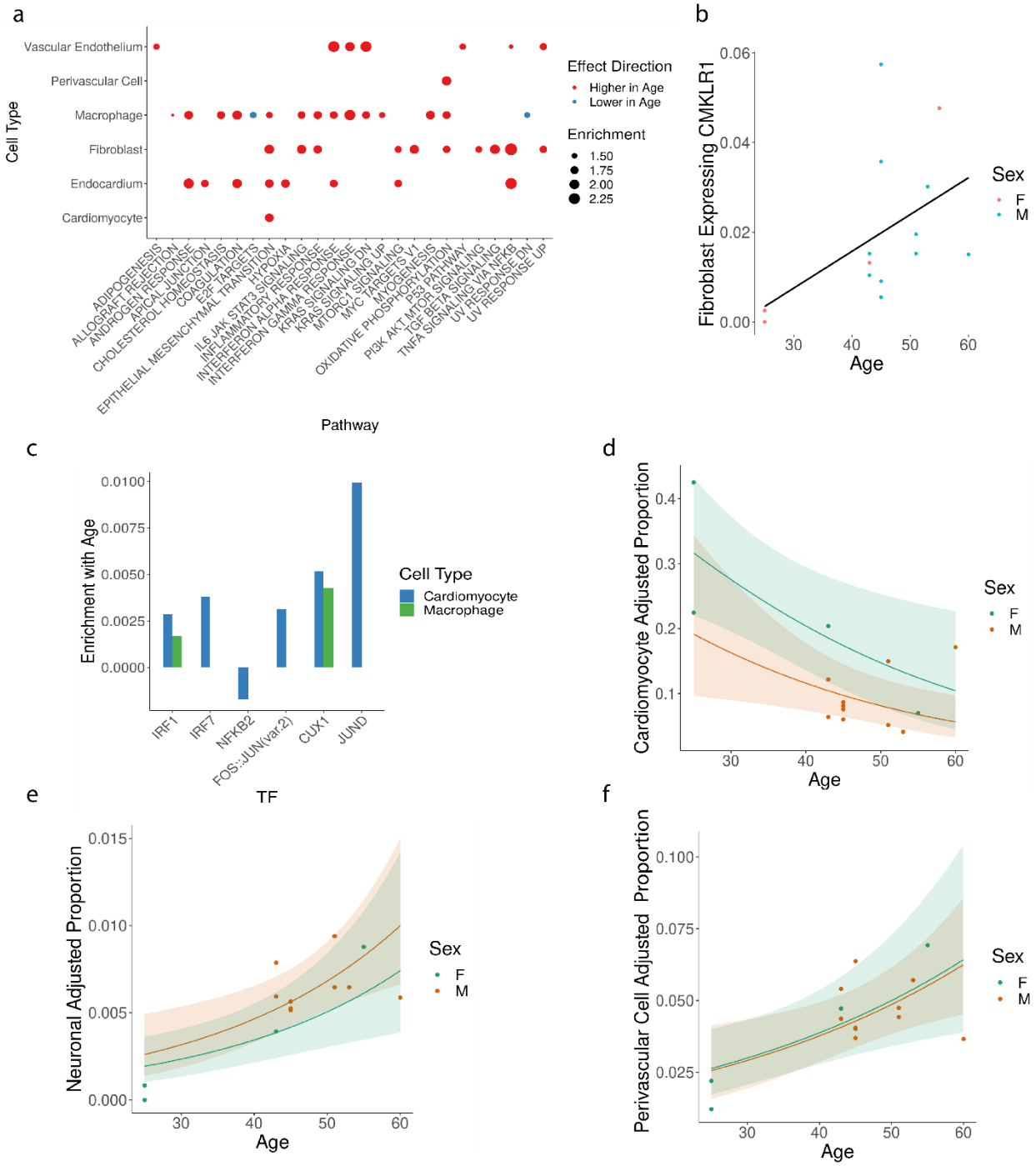

Figure 3-3: Alterations by age in the heart. A) Enrichment of biological processes by age in individual cell types. Pathways "Higher in Age" indicates higher-than-expected expression in older cells after controlling for other covariates, and vice-versa. B) Percent of fibroblasts that are positive for CMKLR1 expression, by age of donor. C) Enrichment for counts of TF motifs in the accessible peaks of cell types as a function of age. Bars shown correspond to statistically significant coefficients (FDR = .1). D) Age of sample donor compared to the adjusted proportion (modified by the effect of anatomical site on proportion in a beta-binomial model) of all cells in that sample that were annotated as cardiomyocytes. Trend line shown based on predicted proportion of male (orange) and female (green) samples over a range of ages. Error range shows +/- 2 standard errors.

additionally see increases in JUND motifs in aged cardiomyocytes (Figure 3.3C) a factor described as a hallmark of senescence in aging (M. J. Zhang et al. 2021). Thus, we observe cell-type specific alterations in TF motif accessibility with age for a handful of regulators associated with senescence. Future single-cell analyses - along with a deeper understanding of senescence markers - will clarify the extent and cell-type specificity of senescence in the aging human heart.

As alterations in cell type abundance - such as declines in cardiomyocyte numbers (Kajstura et al. 2010; Olivetti et al. 1991) and loss of cardiac stem cells (Capogrossi 2004) - are a characteristic of cardiac aging, we looked for alterations in cell type proportions by donor age. As expected, we observe a statistically significant decrease in the proportion of cardiomyocytes captured by donor age (Figure 3.3D). We additionally observe alterations in other cell types, such as increases in the proportion of neuronal and perivascular cells with age (Figure 3.3E,F). Our observations suggest that changes in cellular proportions over time may be widespread. As cardiac function entails complex inter-cell interactions (Hulsmans et al. 2017; Hall et al. 2021), such alterations may manifest in significant functional changes through disrupting intercellular signaling in addition to altering functions carried out by an individual type.

3.4.4 *Contrasting TF motif enrichments identify putative adult- and fetal-specific regulators*

To study the global relationship between TF activity in adult cells versus their embryonic counterparts, we used a regression approach to identify TF motifs enriched in accessible chromatin of specific cell types (see “Methods” section) and compared the enriched motifs in adult cell types against corresponding fetal cell types (Domcke et al. 2020). Motif accessibility in most fetal cell types was largely maintained in the corresponding adult types (Table 3.2), though concordance was weaker for less abundant cell types such as adipocytes, or cell types for which our matching

approach was less confident (e.g. fetal perivascular cells and adult smooth muscle cells) (Figure 3.4).

We next sought adult- or fetal-specific regulatory factors by looking for outliers whose enrichment was markedly higher in one developmental context versus the other. For some cell types, we see few if any obvious discrepancies in TF enrichments between fetal and adult cells.

Cell Type	Embryonic/Adult Correlation	pValue
Cardiomyocyte	0.71	7.15E-88
Vascular Endothelium	0.50	8.98E-35
Endocardium	-0.28	9.95E-10
Macrophage	0.66	1.18E-71
Perivascular	0.04	0.465942
Fibroblast	0.44	1.39E-26
Adipocytes	0.07	0.160144
Neuronal	0.43	3.62E-25
T Cell	0.51	3.15E-37

Table 3-2: **Correlation between corresponding fetal and adult motif enrichments in accessible chromatin.**

For example, TF enrichments in cardiomyocytes are highly correlated between our analysis and corresponding results from fetal data including MEF2 regulators being exceptionally enriched in accessible fetal and adult cardiomyocyte chromatin (Figure 3.4A). This is consistent with MEF2 TFs playing a crucial role in both cardiomyocyte differentiation and maintenance (Desjardins and Naya 2016).

Although the overwhelming majority of TF motifs were similarly enriched in both fetal and adult cells, apparent differences in motif enrichments occur between fetal and adult cell types. For example, while TF enrichment magnitudes are broadly correlated between adult heart and fetal neurons (Pearson correlation = .42, p =3.6e-25. Figure 3.4B), a handful of factors show higher enrichment in adult neurons. Cardiac neurons play a pivotal role in regulating cardiac electrical and mechanical activity through a combination of intrinsic and central-nervous-system interfacing

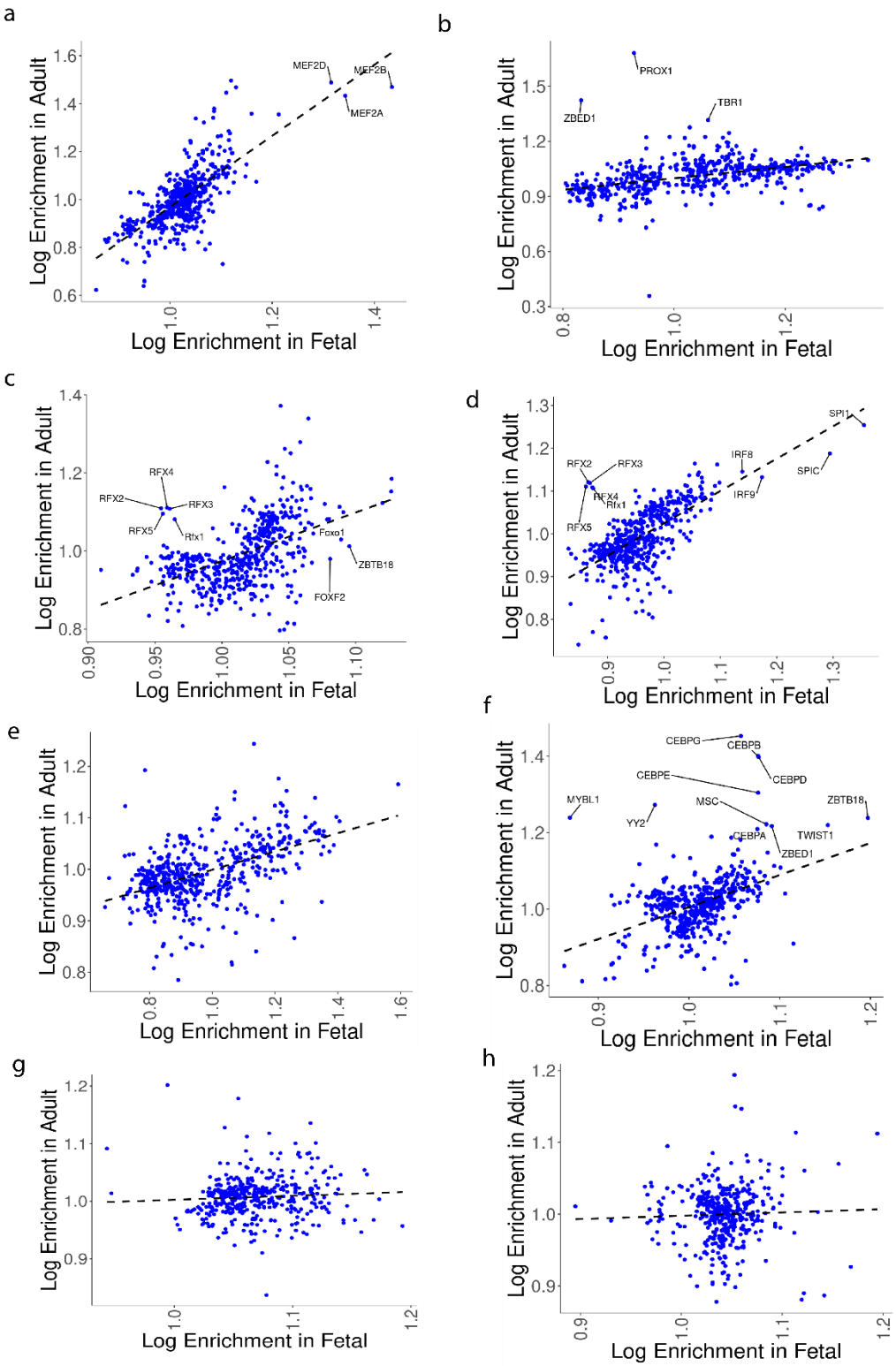

Figure 3-4: **Enrichment of TF motifs in the accessible peaks of fetal or adult sn ATAC-Seq.** A) Enrichments in cardiomyocytes B) Enrichments in cardiac neurons. C) Enrichments in vascular endothelial cells. Enrichments are shown for TFs that were statistically enriched in one or both of fetal or adult analyses (FDR = .1). D) Enrichments in macrophages (adult) versus myeloid lineage cells (fetal). E) Enrichments in T Cells F) Enrichments in fibroblasts (adult) versus stromal cells (fetal). G) Enrichments in perivascular cells (adult) versus smooth muscle cells (fetal). H) Enrichments in adipocytes

interactions (Fedele and Brand 2020), while cardiac neuron dysfunction is central in cardiac arrhythmias (Shen and Zipes 2014; Jungen et al. 2017). In adult cardiac neurons we see notable enrichment for PROX1, ZBED1, and TBR1 motifs in accessible chromatin, contrasting with minimal enrichment (or depletion) of those motifs in fetal cardiac neurons (Figure 3.4B). PROX1 plays a role in cell cycle exit and terminal differentiation of neurons in the central nervous system (Stergiopoulos, Elkouris, and Politis 2014) while TBR1 is essential for neural specification in the developing cortex (Bedogni et al. 2010). ZBED1 plays roles in suppressing cell division (Jin et al. 2020), but apart from possible interactions between a ZBED1 homologue and a regulator of optic lobe formation in *Drosophila* (Jin et al. 2020), ZBED1 has not been previously characterized as a neural regulatory factor. In contrast to adult-specific motif enrichments, we observe fetal-specific enrichments of factors such as NOTO, a regulator of notochord lineage commitment (Colombier et al. 2020), and RORA, a regulator of CNS development (Gold et al. 2003). Altogether, our results show that while some factors are shared between fetal and adult cardiac neurons, others may be developmentally specific.

In fetal vascular endothelial cells, we see enrichment specifically in fetal cells for known vasculature regulators FOXO1 and FOXF, both of which cause severe vascular remodeling defects and embryonic lethality upon knockout in mice (De Val and Black 2009) (Fig. 4C). In addition, we see a similar level of enrichment in accessible fetal chromatin for ZBTB18 in contrast to minimal enrichment in adult vascular endothelium (Figure 3.4C). In the opposite direction, we see adult-specific enrichment for five RFX factor motifs (Figure 3.4C) in adult vascular endothelium. Interestingly, we see adult-specific enrichment for these motifs in a comparison of adult versus fetal macrophages as well (Figure 3.4D). Both of these cells types play crucial roles in vascular dysfunction (Shirai et al. 2015), while RFX factors are correlated with epigenetic changes in

hypertension patients (Reyes-Palomares et al. 2020) and RFX1 indirectly reduces monocyte recruitment in atherosclerosis (Jia et al. 2019). Given motif similarities between RFX factors, further work will be particularly important to understand the role of particular RFX TFs in endothelial and macrophage function. For now, our work raises the potential for adult-specific roles for RFX factors in cardiac endothelium and macrophages joining RFX factors' previously characterized pleiotropic roles (Sugiaman-Trapman et al. 2018).

Overall, our comparison of cell-type specific fetal and adult motif enrichments tells us two things. First, we reproduce the observed correlation across various tissues between corresponding adult and fetal cell types' chromatin in terms of accessible motifs (K. Zhang et al. 2021). Second, motifs that are not correlated between fetal and adult cells identify candidates for developmental stage-specific regulators in cardiac cell types.

3.4.5 *ATAC-Seq links distal sites that improve models of RNA expression*

Characterizing the regulatory roles of noncoding DNA sequences is a pressing challenge in human genetics (Gusev et al. 2014). Although a handful of distal elements with significant functional roles in the heart have been characterized (Alexanian et al. 2021) and genome-wide maps of cis-regulatory elements have recently been published (Hocker et al. 2021), we lack a genome-scale quantitative model of how noncoding sequences drives gene regulation. One approach to linking sequence to transcription has been to train computational models that predict each gene's expression based on nearby sequences and/or epigenetic features (González, Setty, and Leslie 2015; Osmanbeyoglu et al. 2019; Duren et al. 2017; Cheng et al. 2011; Dong et al. 2012). For example, we previously predicted gene expression based on sequence motifs in the accessible chromatin of differentiating myoblasts and found that simple transcription factor motif

presence/absence explained ~37% of transcriptional changes during differentiation (Pliner et al. 2018). Strikingly, information from distal DNA sequences dramatically improved accuracy compared to a model that used only the promoter sequence, suggesting that much of the information needed to encode the cell-state specific expression resides in distal sequences (Pliner et al. 2018). However, the extent to which such models generalize beyond simple *in vitro* systems to multiple *in vivo* human cell types is not clear.

To assess the potential of each cell type's accessible chromatin to predict its transcriptome, we modeled cell-type-specific average gene expression based on promoter sequence alone or in combination with distal sites linked by ATAC-Seq information, as in our previous work (Pliner et al. 2018). We defined hyperparameters for these cell-type specific expression models using a training set, holding aside two separate validation and test sets to measure model performance. Protein coding genes were split into train/validation/test sets at the level of whole chromosomes (see methods) in proportions of approximately 80% train/10% validation/10% test.

We found that in models using only promoter sequence the best average predictive accuracy occurred when the promoter region covered 2000 bases upstream and 1000 bases downstream of a TSS (Figure 3.5A), while use of larger or smaller regions led to inferior accuracy. After finalizing all hyperparameters (Figure 3.5B), two models were fit for each cell type: One used motifs absence/presence in a promoter region only as input features, while a second model used motifs found in promoters or distal DNA sites. In every cell type, models fit using proximal and distal sequence outperformed the corresponding model using promoter motifs alone, for several cell types by nearly 2-fold (Figure 3.5C). Notably, this effect does not appear to be due simply to adding additional arbitrary sequence as use of an even larger promoter region reduced model accuracy (Figure 3.5A).

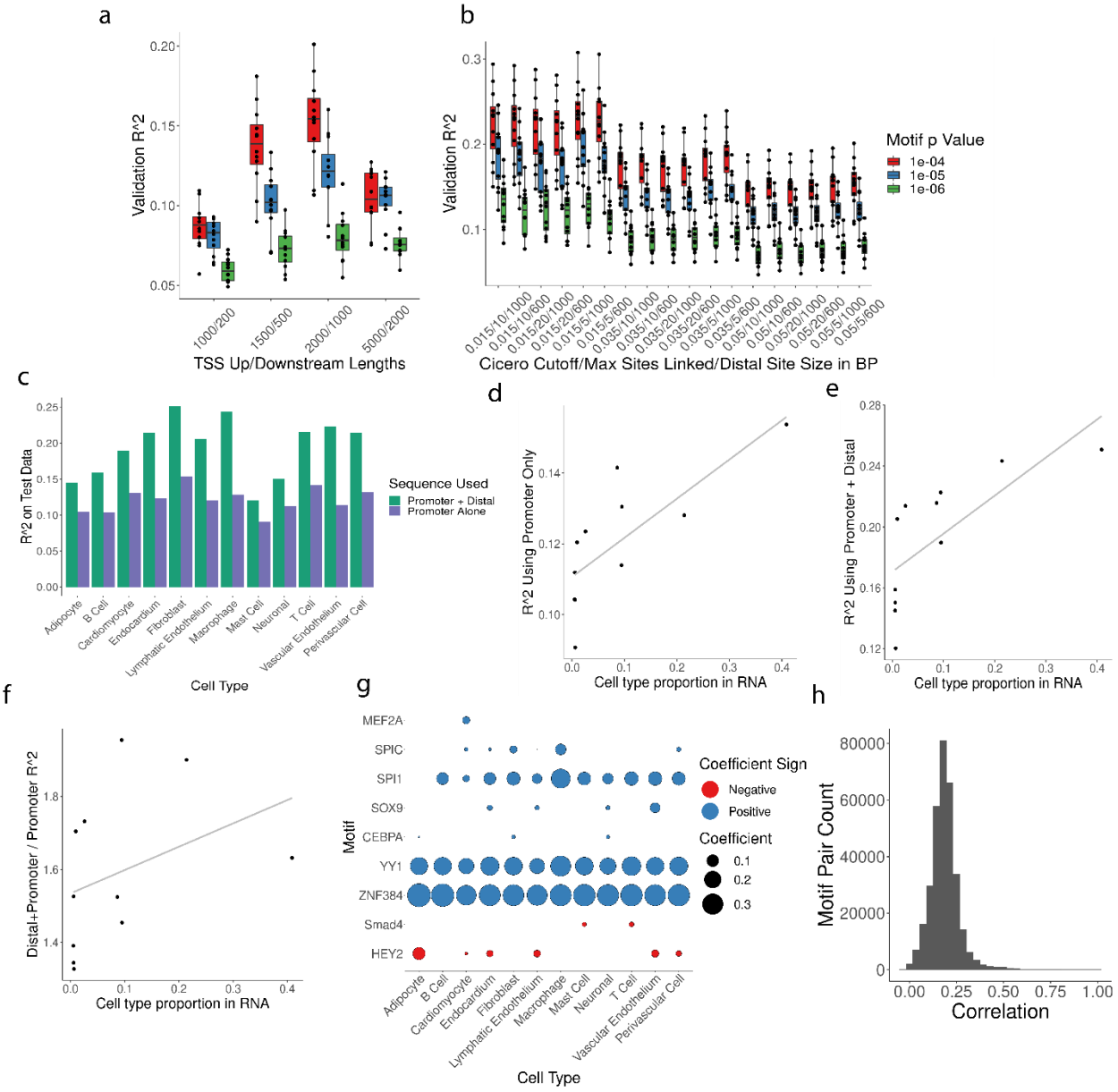

Figure 3-5: Predictive models of RNA expression. A) R^2 for cell type-specific models (one point = one cell type) using varied p value cutoffs for calling motif presence in FIMO (color) and upstream/downstream regions (in bp) with respect to gene TSS. B) Test set R^2 for different cell types. R^2 is listed for models that were trained using motifs found near gene TSS (in purple) or in either TSS or linked distal sites (in green). C) Model accuracy on test data for models using promoter and distal motifs (x-axis) and the proportion of that cell type in the sn RNA-Seq data (y-axis). D) Cell type proportion of total vs. R^2 of model using on promoter motifs E) Proportion vs. R^2 using motifs in promoter or distal sequence F) Proportion vs. the R^2 ratio using promoter and distal sequence or promoter only. G) Magnitude and direction of motif coefficients in final models fit using distal and proximal sequence. H) Distribution of inter-motif correlations across all motif pairs.

Because models for some cell types were more accurate than others (Figure 3.5C), we investigated if model performance was related to the abundance of the cell types. We found that cell type abundance - as quantified by a type's proportion of total cells in RNA-Seq data - was related to accuracy for each respective cell type for models trained on promoter sequence alone (Figure 3.5D) or distal sequence plus promoter sequence (Figure 3.5E). Additionally, models for abundant cell types were markedly improved by including distal information, whereas models for less abundant cell types benefitted less (Figure 3.5F) which suggests that collecting further data would improve our models of gene expression. For abundant cell types, a simple model that predicted each gene's expression based on whether or not each motif was present in the accessible chromatin nearby was able to account for ~20-25% of expression variation at the level of pseudo-bulked transcriptomes, and that around a third of that amount was due to the inclusion of distal motif information. These results demonstrate that for all cell types of the human heart for which we trained models, distal noncoding DNA improves the accuracy of predicted expression.

We next scrutinized the features used by our models to identify the specific sequences that define cell-type specific gene expression. As expected, many motifs that predict expression were also detected as enriched in accessible chromatin (Figure 3.5G). Examples include motifs for SPIC and SPI1 in the macrophage model, MEF2A in the cardiomyocyte model, SOX9 in the vascular endothelial and endocardium models, and CEBPA in the fibroblast model (Figure 3.5G). The models also explicitly identified motifs that are predictive of reduced expression, identifying putative repressors. For example, motifs for HEY2 - which contributes to cardiomyocyte specification in development (Ihara et al. 2020) - were inversely associated with expression in cardiomyocytes and other cell types (Figure 3.5G). This result agrees with the factor's known role as a transcriptional repressor (Xiang et al. 2006), and would be consistent with HEY2 playing a

role in other cell types apart from its characterized function in cardiomyocytes (Ihara et al. 2020). Similarly, the models for T cells and mast cells all captured an inverse relationship between expression and SMAD4 (Figure 3.5G). TGFB signaling via nuclear translocation of SMAD4 is highly cell-type specific, driving a broadly immune-suppressive role (Batlle and Massagué 2019) with SMAD4 alternately acting as a transcriptional repressor or activator depending on cellular context (Wotton et al. 1999). In addition, we observed cases where a particular TF motif was utilized by models across cell types, such as ZNF384 and YY1 motifs leading to increased expression predictions in all cell types (Figure 3.5G). Such relationships are difficult if not impossible to detect in a testing strategy looking for enrichments in a cell type compared to others. In summary, predictive models identify TFs that play cell type-specific roles, TFs related to expression across many cell types, and assign an explicit direction of effect on transcription.

3.5 DISCUSSION

We generated a resource of snATAC- and snRNA-Seq from multiple donors and utilized a state-of-the-art regression approach to find alterations by sex and age. We additionally studied regulatory roles of transcription factors in distinct cell populations through enrichment analyses and predictive expression models. Performing single-cell or single-nucleus RNA analysis in solid human tissue is difficult, and our dataset represents one of only a small handful of studies covering the healthy human heart at single-cell resolution from multiple donors (Litviňuková et al. 2020; Tucker et al. 2020; Hocker et al. 2021; K. Zhang et al. 2021). Our data allows us to characterize alterations at multiple levels, such as finding sex-specific TGFB-driven transcription (Figure 3.2B) accompanied by altered epigenetic signatures of TGFB effectors and regulators (Figure 3.2F). Cell-type specific changes - such as statistically significant TNFA activation hallmarks in male T cells but no other cell types (Figure 3.2B) - and altered cell type abundances (Figure 3.2E and

3.3E) highlight the key role of single-cell resolution for understanding alterations by donor traits. Such cell type-specific changes could contribute to why a recent analysis of sex-dependent bulk transcriptional changes found most sex-specific changes to be of small magnitude and detected a minimal number of statistically significant pathway alterations in heart samples (Oliva et al. 2020). Our results indicate that single cell libraries from a tractable number of donors can identify diverse alterations in cell type proportions, transcriptional differences, and epigenetic programs varying by donor traits.

In addition to facilitating tests of sex- and age-dependent differences across cell types, our dataset also allowed us to quantify the level of transcriptional variation that can be accounted for using a simple, binary TF motif-based linear model when using or excluding TSS-distal sequence (Figure 3.5C). Our results underscore the importance of distal regulatory information in determining gene expression levels (Duren et al. 2017; Pliner et al. 2018; Zeng, Wang, and Jiang 2020) and reaffirm the observation that - at least in a simple modeling framework - simple models using TF motifs as input features explains only a minority of total variation in RNA levels (González, Setty, and Leslie 2015; Osmanbeyoglu et al. 2019). Furthermore, our models identify regulatory factors through a method that complements widely-used tests (Hocker et al. 2021; K. Zhang et al. 2021; Domcke et al. 2020) based on motif presence in accessible peaks (Figure 3.1F). In total, our predictive models of expression extend existing analyses of cell-type specific regulatory programs in the adult human heart (Hocker et al. 2021; K. Zhang et al. 2021) and establish a baseline level of accuracy for comparison against more complex methods (Avsec et al. 2021; Agarwal and Shendure 2020; J. Zhou et al. 2018) utilizing larger datasets and less interpretable models.

Our analyses come with several caveats, most notably due to the limited number and diversity of samples analyzed. While we find statistically significant alterations by age and sex within the samples of our dataset, the limited number of individuals studied precludes generalizing those differences to the population at large. Considerably larger datasets will be necessary to find generalizable differences while accounting for confounding variables like donor medical history and disease status, dissect finer-grained effects like sex-specific differences before or after menopause onset (Kessler et al. 2019), and provide statistical power to detect biologically meaningful changes of small magnitude. In addition, analyses of new datasets - or meta-analysis of existing ones - will be required to study differences occurring in cell populations like atrial cardiomyocytes (Litviňuková et al. 2020; Tucker et al. 2020) that we did not study because of the limited anatomical coverage of our samples. In our predictive models of expression, due to limited training inputs and a preference for easily interpretable models we employed a simple linear model based on motif presence/absence. Future work in more complex - and likely more accurate - models like deep neural networks (Avsec et al. 2021; Agarwal and Shendure 2020; J. Zhou et al. 2018) will be required to understand the relative value of distal DNA sequence and of particular sequence features/motifs in alternate prediction frameworks. It will be particularly interesting to test the reproducibility of cell-type-specific TF roles found via our linear models (Figure 3.5G), given inherent co-occurrence of motifs in the genome (Figure 3.5H). As modeling approaches that use regularization to select a minimal set of predictors (e.g. LASSO and elastic net regression) will often pick only one member of a motif "family", motif/expression relationships in particular cell types will require validation in further modeling approaches.

Our findings raise a number of natural avenues for further research. First, analyses of larger datasets of single-cell heart tissue - either newly generated, or via meta-analyses of existing

datasets - will be crucial for identifying sex- and age-dependent variation that generalizes to the population at large, and to extend analysis to cover other important patient covariates or interactions. Relatedly, similar analyses in other organs can study the extent to which sex- and age-dependent variation observed in cell types of the heart is reproduced in cell types of other human tissues. Furthermore, perturbational work in animal or culture models will play a crucial role in studying differences detected in large-scale, single-cell analyses by avoiding limitations inherent to observational designs. Our identification of biological processes such as metabolic shifts, TGFβ signaling, and inflammation also raise the prospect that larger analyses may detect alterations by sex or age that are already of clinical interest in cardiac disease (Lopatin 2015; Parichatikanond et al. 2020; Kosmas et al. 2019) and raise the possibility of personalizing therapies by patient traits. Reassuringly, our identification of numerous statistically significant differences in both RNA- and ATAC-seq data using a modest number of donors bodes well for the ability of future analyses to identify molecular covariates of patient traits within datasets that can feasibly be generated with current technology.

Chapter 4. APPLICATION OF SCI RNA-SEQ TO MOUSE MODELS OF PULMONARY ALVEOLAR PROTEINOSIS AND SILICOSIS

4.1 INTRODUCTION

As tissues of multicellular organisms involve complex molecular interactions within and between cells, disease in complex organisms inevitably involves alterations spanning a spectrum of molecular pathways across specialized cell types. Due to the variety of changes that characterize tissue perturbation and the innate functional specialization of cells, single-cell RNA-Seq methods have been widely used in recent years to profile alterations occurring in tissue disease in a cell-type-resolved manner (Koenig et al. 2022; Reichart et al. 2022; Chaffin et al. 2022; Adams et al. 2020; Puvogel et al. 2022; Ruzicka et al. 2020; Conway et al. 2020). During my work in the Trapnell lab, I worked on the data generation portion of two projects in that trend, where our lab generated and analyzed single-nucleus RNA-Seq datasets from mouse models of Pulmonary Alveolar Proteinosis and Silicosis. Though distinct diseases, we sought to leverage shared single-nucleus RNA sequencing methods and analytical approaches to learn about the molecular hallmarks of these diseases, as well as set the stage for larger analyses of lung disease attributes that are shared across various forms of lung dysfunction.

Pulmonary alveolar proteinosis (PAP) is a rare lung disease in which a buildup of alveolar surfactant leads to impaired gas exchange and lung failure (B. C. Trapnell et al. 2019). PAP occurs due to distinct causes, with disease in “primary PAP” caused by disruption of granulocyte-macrophage colony-stimulating (GM-CSF) signaling. Disruption of GM-CSF leads to a loss of differentiated alveolar macrophages and subsequent accumulation of surfactants and cholesterol due to loss of the surfactant degradation activity by alveolar macrophages (B. C.

Trapnell et al. 2019). GM-CSF signaling can be ablated by autoantibodies against the GM-CSF ligand, genetic loss of GM-CSF ligand, or genetic loss of one of two receptors (GM-CSFRA or GM-CSSFRB) (Borie et al. 2011).

In addition to genetic diseases such as primary PAP, another important area of lung dysfunction involves inhalation of damaging particulates. For example, silicosis is a disease caused by accumulation of silica particulates in the lungs. Silica deposits that cannot be degraded lead to reduced lung function and eventual lung failure characterized by inflammation and fibrosis (Cassel et al. 2008; Leung, Yu, and Chen 2012; Wagner 1997). Minimal therapeutic options currently exist for silicosis, though current studies are underway to assess the utility of anti-fibrotic and anti-inflammatory small molecule therapies (T. Li et al. 2022). Despite the potential value of new treatments, uncertainty surrounding molecular alterations at play in a silicosis-afflicted lung – much less and understanding of the essential causal links underpinning lung dysfunction – makes principled therapeutic development challenging.

There is significant uncertainty in silicosis and PAP as to how the various cell types of the lung change in disease. In PAP, outstanding uncertainties include aspects such as whether semi-differentiated alveolar macrophage precursors persist in the lung, how alveolar epithelial cells are altered in the presence of pathological surfactant accumulation, and what molecular signatures are generated by distinct cell populations to drive inflammation (B. C. Trapnell et al. 2019). In silicosis, areas of interest include the heterogeneity of macrophage states in the face of persistent particulates, the full repertoire molecular programs utilized by epithelial cells in the injured lung, and the molecular basis of the transition between inflammation and fibrosis (Barnes et al. 2019).

To better understand the cell-type-resolved alterations in lungs affected by PAP or silicosis, we sought to generate sci RNA-Seq datasets for mouse models of PAP and silicosis. To generate this data, we revisited the original method used to prepare nuclei for sci RNA-Seq (Cao et al. 2017) to be compatible with nuclei from solid, frozen tissue. Through a series of experiments, we optimized a method for extracting nuclei from frozen tissue such that RNA was not degraded and nuclei were fixed in a way that was compatible with sci RNA-Seq workflows. With this method, we generated single nucleus RNA-Seq datasets from mouse models of silicosis and PAP, revealing distinct alterations in lung cell types as a consequence of disease.

4.2 METHODS

4.2.1 *Isolation and fixation of nuclei from mouse tissue for sci RNA-Seq*

For generation of all mouse lung datasets, snap-frozen mouse lung tissue (93-223 mg) was dissociated using a Gentle MACS tissue dissociator using "C" dissociation tubes in 5 mL of ice-cold lysis/fixation buffer (10 mM NaCl, 10 mM sodium phosphate pH 7.2, 3 mM MgCl2, 5% glutaraldehyde, 10 mM vanadyl ribonucleoside complex, .1% Triton X-100, 1% diethyl pyrocarbonate, .00015% polyvinyl sulfonic acid (Sigma Cat. 278424)). Tissue was dissociated using the "Mouse Spleen 1" program for 60 seconds, then filtered using a 70 uM cell strainer. The strainer was washed with an additional 5 mL lysis/fixation buffer, then nuclei were fixed at 4C for 15 minutes. Nuclei were pelleted by centrifugation at 500 RCF, 4C, 8 minutes. Supernatant was discarded and nuclei were resuspended in 1 mL nuclei suspension buffer (10 mM Tris HCl pH 7.4, 10 mM NaCl, 3 mM MgCl2). Nuclei were filtered through a 30 uM strainer, then pelleted by centrifugation at 500 RCF, 4C, 5 minutes. Supernatant was discarded and nuclei were resuspended in 500 uL of nuclei suspension buffer and pelleted again at 500

RCF, 4C, 5 minutes. After discarding the supernatant, the nuclei pellet was resuspended in 210 uL of nuclei suspension buffer. 2 aliquots of 100 uL nuclei suspension were snap-frozen in liquid nitrogen and stored in liquid nitrogen storage for sci RNA-Seq library preparation.

Variations were used in experiments optimizing the above, final protocol. In experiments testing RNase inhibitor use (Figure 4.1A), the use or non-use of vanadyl ribonucleoside or DEPC content in the lysis buffer is indicated by the axis label for each sample. In experiments using different dissociation methods (Figure 4.1A and B), the dissociation method is noted on the axis label. “Mince” refers to mincing with a razor blade, “Hammer” refers to powdering frozen tissue while foil-wrapped on dry ice, and “MACS” refers to dissociation with a Gentle MACS dissociator as described above. In experiments varying fixation (Figure 4.1C), the 5% glutaraldehyde content described above was varied to some level of paraformaldehyde or glutaraldehyde as specified by axis labels. In the experiment varying FACS sorting (Figure 4.1E-G), following reverse transcription in a 2-level sci RNA-Seq preparation some nuclei were FACS sorted into wells for tagmentation and PCR using FACS, while others were counted and diluted to appropriate concentrations before distributing to tagmentation/PCR wells by pipette.

4.2.2 *Sci RNA-Seq Library Generation*

Single cell RNA-Seq data was prepared using the 2-level workflow for sci RNA-Seq (Cao et al. 2017). The protocol was modified to use the following RT incubation temperatures, instead of a 55C/5 minute incubation: 2 minutes at each of 4C, 10C, 20C, 30C, 40C, 50C followed by 10 minutes at 53C and 15 minutes at 55C. Each sample was distributed into 48 wells (1/2 an RT plate) of uniquely indexed RT reactions, then

Samples were prepared using 6 plates of RT indices (for 576 indices used) divided evenly between the 12 samples. 150 nuclei in total were sorted into each well of the 4 PCR plates that were prepared. Libraries were sequenced using an Illumina Nextseq 550.

4.2.3 *Mouse Model for PAP*

Mice for use in this study included four genotypes: wild type, as well as double knockouts for Csf2, Csf2rb, and Csf2ra. Wild type mice are C57Bl/6, and mouse KO models are described for Csf2-/- at (Stanley et al. 1994), for Csf2rb-/- at (Robb et al. 1995), and for Csf2ra-/- at (Shima et al., n.d.). Mice were sacrificed at 12-13 weeks of age, then whole lungs were removed and snap-frozen in liquid nitrogen for later dissociation.

Note: Analysis of PAP RNA-Seq datasets described here and in Chapter 4.4.2 was performed by Dr. Claire Williams. Methods descriptions are adapted with minimal modification from a manuscript under preparation describing these experiments.

4.2.4 *Sci RNA-Seq Analysis for PAP Experiments*

Raw sequencing data was processed using the Brotman Baty Institute pipeline (Pliner, Gogate, and Ewing n.d.). Briefly, the pipeline processes and aligns reads to the Mm9 mouse reference genome, merges duplicate reads by UMI sequence, and assigns reads to individual cells based on matching index sequences indicating unique PCR wells and RT wells.

RNA-Seq data was analyzed using Monocle3 (Cao et al. 2019). Doublets were identified and removed using Scrublet to discard cells with Scrublet Scores < 0.2 (Wolock, Lopez, and Klein 2019), and low-quality cells were filtered based on low UMI (<100 UMI per cell) or high percentage of reads mapping to mitochondria (>10%). Cells were subject to dimensionality

reduction using UMAP (McInnes, Healy, and Melville 2018) using default parameters as implemented in the Monocle3 reduce_dimension function (metric = “cosine”, min_dist = .1, n_neighbors = 15, nn_method = “annoy”, top 100 components from PCA as input features) after correcting for batch effects and genotype using mutual nearest neighbors (Haghverdi et al. 2018) and regressing out log10(UMI) using the align_cds function in Monocle3, which fits a linear model for the relationship between cells’ UMIs and coordinates in PCA space then subtracts the effect of UMI on those coordinates using LIMMA (Ritchie et al. 2015). Cells were clustered after UMAP embedding using Leiden clustering (Traag, Waltman, and van Eck 2019) within the cluster_cells function of Monocle3, and cell types were assigned manually based on expression of literature-derived marker genes (Han et al. 2018; Angelidis et al. 2019).

Differential expression tests used a negative binomial mixed effect model fit using lme4 (“Linear Mixed-Effects Models Using ‘Eigen’ and S4 [R Package Lme4 Version 1.1-28]” 2022) implemented within Monocle3. Individual was treated as a random effect, while genotype and log10(UMI) were treated as fixed effects.

Changes in cell type proportion were tested using a beta-binomial model (B. D. Martin, Witten, and Willis 2020) comparing across genotypes. For tests of enrichment/depletion we use a Poisson-Log Normal (PLN), a multivariate mixed generalized linear model with a Poisson distribution as its underlying statistical framework. This convenient framework allows you to both 1) perform multivariate statistical regression to describe how genotype relates to the relative abundances of each cell state and 2) describe how all pairs of states co-vary as a parsimonious network of partial correlations. We hypothesize that PLN network models will accurately quantify shifts in the distribution of cells over molecular states following genetic perturbations.

For pseudotime analysis (C. Trapnell et al. 2014), cells from terminal populations were used to generate a new UMAP embedding. Within that embedding we generated a pseudotime trajectory using the learn_graph function in Monocle3.

4.2.5 *Mouse models of silicosis*

C57BL/6J mice were obtained from Jackson Laboratory (Bar Harbor, ME). Sprague-Dawley rats were obtained from Charles River Laboratory (Wilmington, MA). For terminal experiments, mice or rats were euthanized by i.p. injection of Euthasol (Henry Schein, Melville, NY). All animals were maintained in a specific pathogen–free facility and were handled according to a University of Cincinnati Institutional Animal Care and Use Committee –approved protocol and National Institutes of Health guidelines.

Silica particles (Sigma Aldrich, St. Louis, MO, particle size: 80% between 1 and 5 μm) were boiled in 1N HCl for 1 h, washed with dH_2O, and dried at 100°C. The particles were then heat sterilized at 200°C for 2 h and suspended in sterile saline. The endotoxin content in the silica particles was <1.0 pg/μg of silica as determined using the LAL Chromogenic Endotoxin Quantitation Kit (Thermo Scientific, Rockford, IL) according to the manufacturer's instructions. For i.t. silica administration, C57BL/6J mice were anesthetized with isoflurane and suspended by their incisors in the supine position on a procedure board at a 45°C angle. The glottis was visualized by retraction of the tongue and illuminated with a fiberoptic thread. A 22-gauge angiocatheter was advanced into the trachea under direct visualization, and after confirming correct placement by expansion of the thorax upon delivery of air though the catheter, 2.5 or 5 mg of silica in 100 μl of saline was injected into the lung. For oropharyngeal (o.a.) silica administration, each mouse was anesthetized with isoflurane and suspended by a steel wire on a procedure board at a 60°C angle by the incisor teeth. The mouth was opened, the

tongue was pulled forward and 5 mg of silica in 50 µl of saline was placed at the base of the tongue. Once the slurry was aspirated into the lungs with inspiration, the tongue was released.

Note: Analysis of silicosis datasets described here and in Chapter 4.3.3 was performed by Dr. Jennifer Franks. Methods descriptions are adapted with minimal modification from a manuscript under preparation describing these experiments.

4.2.6 *Sci RNA-Seq Analysis for Silicosis Experiments*

Raw sequencing data was processed using the Brotman Baty Institute pipeline (Pliner, Gogate, and Ewing n.d.).

RNA-Seq data was analyzed using Monocle3 v1.2.9 (Cao et al. 2019). Cells were filtered for quality control based on the following thresholds: >100 UMIs, <10% mitochondrial RNA, <0.2 Scrublet (Wolock, Lopez, and Klein 2019) doublet score. For visualization, we performed dimensionality reduction using the first 100 principal components, followed by 2D UMAP projection. We hierarchically annotated the data (into broad and fine cell-states) using Louvain clustering with varying levels of resolution in combination with marker genes identified from literature.

For differential abundance testing, cell numbers were collapsed per sample according to the fine cell state annotation. Cell numbers were normalized to correct for different cell numbers recovered from each sample. A beta-binomial test with Benjamini-Hochberg correction for multiple hypotheses (Benjamini and Hochberg 1995) was used to test for differences in cell state abundance for each timepoint compared to baseline (Day 0). A corrected p-value less than 0.05 was considered significant.

To run gene signature analysis, gene sets were acquired from MSigDB (Subramanian et al. 2005) (Hallmarks Inflammation Gene Set, GO_Osteoclast_differentiation), Aran et al 2017 (Aran, Hu, and Butte 2017)} (M1/M2 signatures), and Wang et al (B. Wang et al. 2020) (Fibrosis signature). For each gene, we translated the human gene set to mouse orthologs using the gorth function from the gprofiler2 package in R. For M1 and M2 analyses, only genes unique to one of the lists was used. A summary score of gene set activation was calculated per cell using the aggregate_gene_expression() function in Monocle3 (Cao et al. 2019). Scores were calculated using log-transformed expression values and were normalized to a scale of -3 to 3.

To test for differences in gene expression, we used a linear mixed effect model with splines. Significant differentially expressed genes were annotated with gene ontology (GO) function terms using g:Profiler. Terms with $p<0.05$ corrected for multiple hypothesis testing with the default g:SCS method were considered significant.

We calculated pseudotime for the interstitial macrophages to capture pathway activation changes associated with macrophage polarization and plasticity. Pseudotime was calculated using learn_graph() in Monocle3 (C. Trapnell et al. 2014) with default parameters except for use_partitions = FALSE. The root node was designated based on the end of the trajectory with highest number of cells collected from Day 0. Genes expressed in at least 100 cells that vary along the pseudotime trajectory were identified using a linear model with natural splines (df = 3). Differentially expressed genes were selected based on $q\_value<0.05$. Significant differentially expressed genes were annotated with gene ontology (GO) function terms using g:Profiler. Terms with $p<0.05$ corrected for multiple hypothesis testing with the default g:SCS method were considered significant.

4.3 RESULTS

4.3.1 *An optimized protocol generates single nucleus sci RNA-Seq data from mouse lungs*

Over a series of experiments we determined four areas of adjustment to the original sci RNA-Seq workflows (Cao et al. 2017, 2019) to generate high-quality single-nucleus datasets from frozen mouse lung.

First, we determined the importance of including effective RNase inhibitors during the initial dissociation of tissue and nuclei extraction. For example, RNA recovery was drastically improved when lysis buffers included diethylpyrocarbonate (DEPC, which inactivates RNases by covalently modifying histidine residues and primary amines) and vanadyl ribonucleoside complex (a transition state analogue of RNA cleavage that acts as a competitive RNase inhibitor) (Figure 4.1A). Inclusion of only the RNase inhibitor content of previous sci RNA-Seq workflows (Cao et al. 2017, 2019), recombinant SuperaseIn, was insufficient to effectively protect RNA during nuclei isolation (Figure 4.1A).

Second, we used glutaraldehyde in place of paraformaldehyde (PFA) to fix nuclei extracted from adult mouse tissue (Figure 4.1B). Use of glutaraldehyde in place of PFA increased UMI recovery per nucleus by ~2 fold.

Third, we determined that dissociation using an automated tissue dissociator provided equal quality data as assessed by median UMI-per-nucleus when compared to a more time-intensive dissociation method (Cao et al. 2019) involving mincing and pressing of tissue.

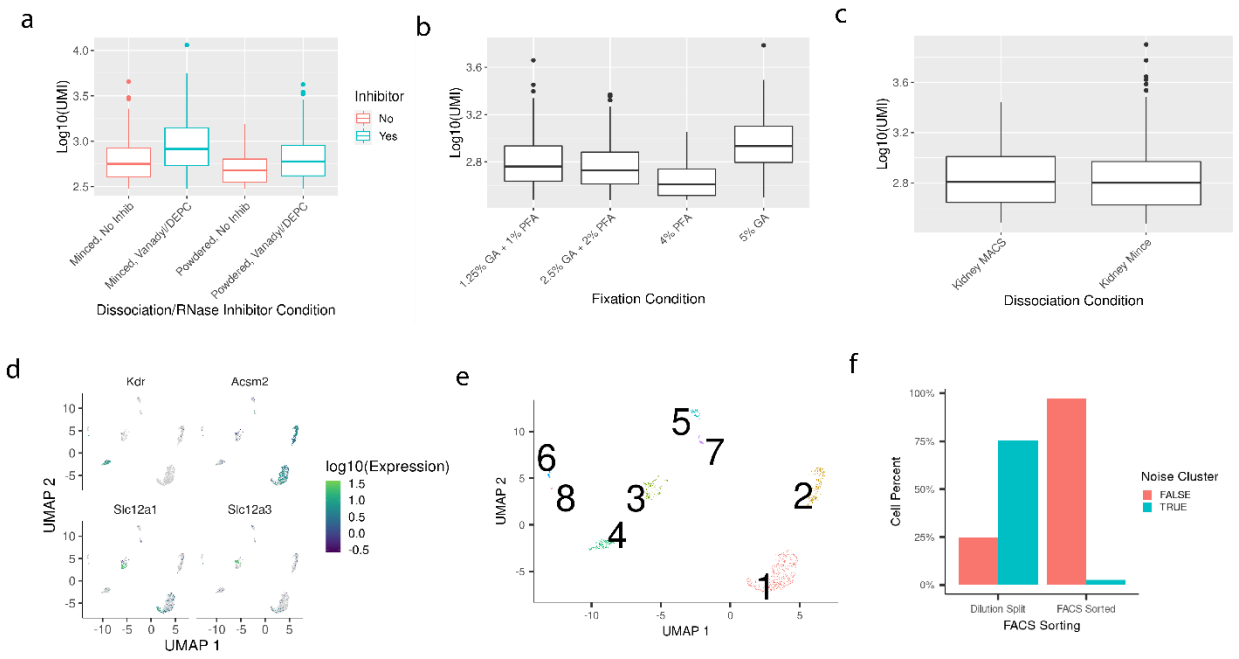

Figure 4-1**: Optimization of nuclei preparation methods for sci RNA-Seq in mouse tissue.** A) Nuclei transcriptomes recovered based on dissociation method and use/non-use of the RNase inhibitors vanadyl ribonucleoside complex and DEPC. B) UMI distributions from nuclei obtained via sci RNA-Seq after tissue dissociation using a GentleMACS Tissue Dissociator ("Kidney MACS") or mincing with a razor blade ("Kidney Mince"). C) UMI distributions for nuclei processed with sci RNA-Seq following fixation with some mixture of paraformaldehyde (PFA), glutaraldehyde (GA), or both D) Faceted UMAP of nuclei obtained from a mouse kidney sci RNA-Seq library. Nuclei are colored by expression of cell-type specific marker genes E) Clusters and cluster numbers for mouse kidney nuclei data F) Percent of nuclei that were either FACS-sorted or not FACS-sorted ("Dilution Split") that belong to the "noise cluster" - cluster 1, in panel e - in a UMAP embedding.

Fourth, we required FACS sorting of nuclei before placement of nuclei into tagmentation/PCR wells during sci RNA-Seq processing. Nuclei prepared by sci RNA-Seq without this additional step exhibited significant levels of RNA cross-contamination, observable as a single "high noise" cluster (Figure 4.1D). In this large cluster, composed mostly of non FACS-sorted nuclei (Cluster 1 in Figure 4.1E), there is expression of markers for multiple distinct cell types (Figure 4.1D). In contrast, FACS-sorted nuclei separate into distinct clusters with cell-type-specific markers showing expression that is specific to individual clusters. We find that most FACS-sorted nuclei end up in clusters outside of the large, multi-marker "noise

cluster" while most unsorted nuclei lie within the noise cluster in a UMAP embedding (Figure 4.1E).

With these alterations determined, we applied our updated sci RNA-Seq workflow to two mouse models of lung disease.

The following section is adapted with minimal modifications from a manuscript under preparation. Writing and analysis presented here was led by Dr. Claire Williams.

4.3.2 *PAP is characterized by alterations in macrophage-like cells*

Patients with primary PAP exhibit severe deficits in lung function, and we sought to define the cellular and molecular changes that drive these functional abnormalities. Mice with mutations in the ligand GM-CSF or its receptors, Csf2, Csf2ra, and Csf2rb, phenocopy key features observed in PAP patients, including buildup of excess surfactant in the lung (B. C. Trapnell et al. 2019). With the goal of characterizing the PAP lung in its entirety, we dissociated whole lungs from twelve-week-old diseased mice harboring each of these three mutations in addition to healthy controls and performed single nucleus RNA-sequencing using two levels of combinatorial indexing on four to eight individuals per genotype (Fig 4.2A). After removing low quality nuclei, we retained 33,576 high quality nuclei across the four conditions for further analysis. Importantly, nuclei were isolated separately from each individual, providing independent biological replicates for each condition.

Nuclei from the healthy lung represented all major cell types expected to be present. Dimensionality reduction with UMAP yielded nine well-separated clusters to which numerous markers of lung cell identity were restricted, enabling us to annotate nine major cell types (Figure 4.2B). In order of abundance in the healthy lungs, the major identified clusters contained

airway epithelial cells (AECs), myeloid cells, endothelial cells, fibroblasts, Type II pneumocytes (AT2), T lymphocytes, Type I pneumoctyes (AT1), B lymphocytes, and mesothelial cells. Sub clustering of these major cell types revealed variegated expression of markers of functional specialization, resulting in putative annotation of 39 distinct subtypes of lung cells (Figure 4.3). The abundant representation of all major lung cell types and the detection of key marker genes within each confirmed that this experiment captured the cellular and molecular diversity expected in the healthy murine lung.

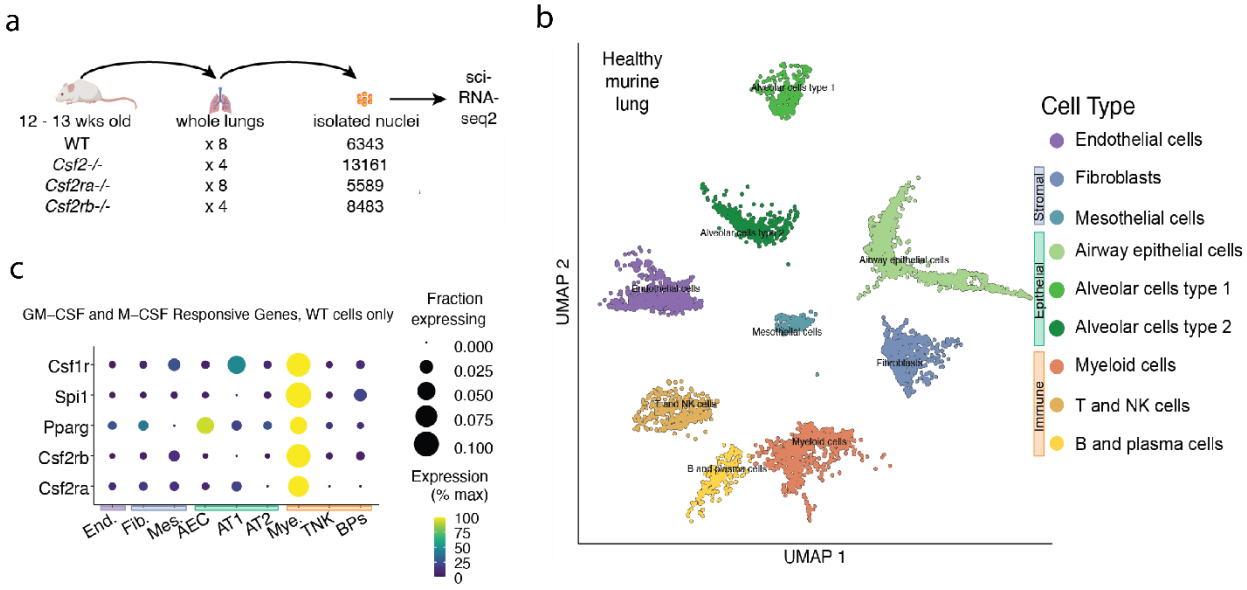

Figure 4-2: **Single nucleus RNA-Seq of PAP genetic mouse models**. A) Experimental design. B) UMAP embedding of wild-type mice data, colored by cell type. C) Expression levels of downstream markers of GM-CSF and M-CSF signaling, by cell type and genotype.

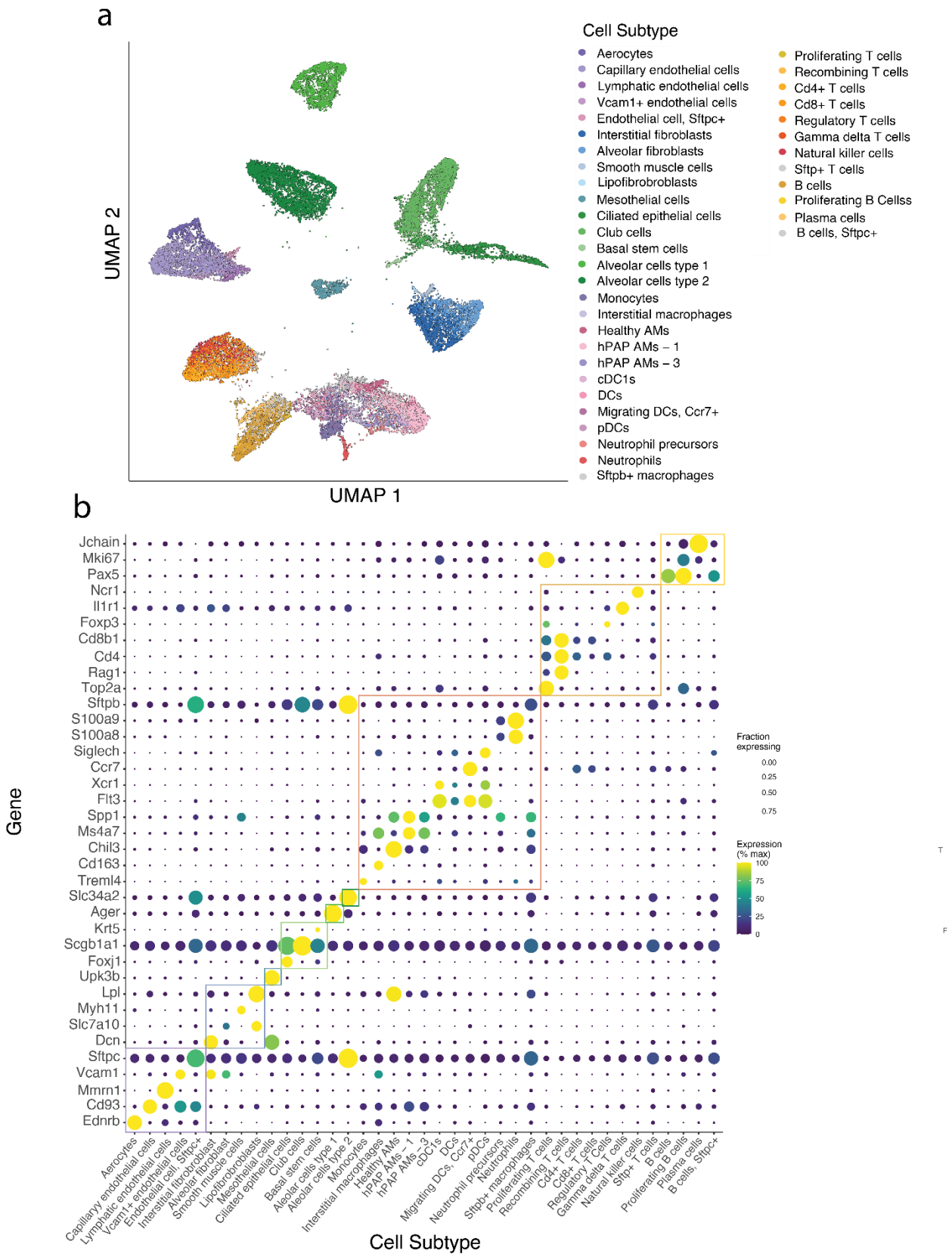

Figure 4-3: Subtypes in RNA-Seq of PAP mouse models. A) UMAP colored by cell subtype. B) Marker expression by cell subtype.

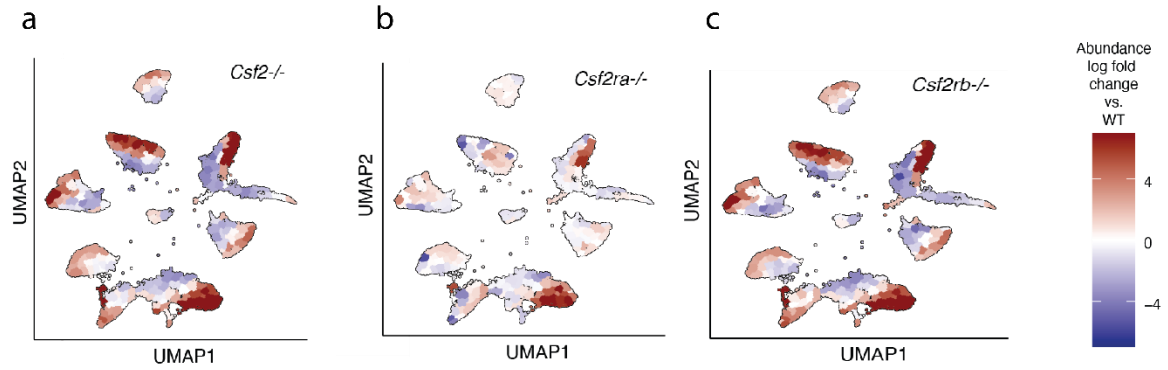

Figure 4-4: **Abundance enrichments and depletions for GM-CSF signaling KO genotypes compared to wild type**. A separate UMAP embedding in shown for A) Csf2-/- mice B) Csf2ra-/- mice C) Csf2rb-/- mice.

GM-CSF signaling is impaired in PAP, so we reasoned that cell types that normally respond to GM-CSF are most likely to be directly impacted by this deficiency. As previously reported (T. Suzuki et al. 2011), myeloid cells expressed high levels of the GM-CSF receptor subunits, Csf2ra and Csf2rb, in addition to downstream signaling components (Figure 4.2C). The key transcription factors Spi1 and Pparg additionally were moderately expressed in B cells and airway epithelial cells, suggesting that GM-CSF deficiency may induce primary changes in the cell states of cell types beyond macrophages. Because there is a buildup of cholesterol rich surfactant in the lung in patients with PAP, we predict there may additionally be indirect effects of GM-CSF deficiency on the cell states of non-GM-CSF responsive cell types.

We next sought to compare healthy lungs against each of the three genetic mouse models of PAP in terms of the cell states and abundances of each major cell type (Figure 4.4A-C). First, the relative enrichment and depletion of cell states in each mutant background as compared to the cell states present in healthy lungs was visualized across UMAP space (See methods). There are subtle shifts in these UMAPs across cell types, suggesting wide-ranging effects of GM-CSF deficiency throughout the lung. The most striking difference was the appearance of abnormal macrophages and a corresponding depletion of healthy alveolar macrophages across all three

PAP models. Additionally, there were regions of elevated occupancy consistently across all three PAP models in B cells, airway epithelial cells, endothelial cells, and fibroblasts. This is consistent with the observation of pulmonary lymphocytosis and fibrosis in patients with PAP (Ebina-Shibuya et al. 2017; Luisetti et al. 2011), and further investigation of these cell proportion alterations is underway given subtleties of interpreting changes in the 2D UMAP space. We observed larger differences in the cell states of two genetic backgrounds, Csf2-/- and Csf2rb-/-, than of the third genetic background, Csf2ra-/-. We speculate that the reduced severity of alpha chain mutations may in part explain the elevated occurrence of this mutation in human patients (Hadchouel et al. 2020).

Myeloid cells are known to be severely affected in patients with PAP and in murine models, with a deficit of mature, healthy alveolar macrophages a key phenotype of the disease (B. C. Trapnell et al. 2019). To more closely examine the myeloid cells in healthy and diseased lungs, we isolated these cells from our dataset and reprojected them into UMAP space, generating a new UMAP space only derived using these cells. By clustering these cells and examining the expression of known markers as well as enriched genes, we identify numerous myeloid cell subtypes including macrophages, monocytes, dendritic cells, and neutrophils (Figure 4.5A). Healthy alveolar macrophages (AMs), expressing key marker genes such as Chil3 and Car4, are largely absent from all three hereditary PAP (hPAP) genotypes (Figure 4.5B). A small number of healthy macrophages are present in Csf2rb-/- potentially suggesting compensation from Csf2rb2 or possible signaling through Csf2ra at higher GM-CSF concentrations in these mice.

In the diseased lungs, there is a compensatory increase in at least two disease-enriched macrophage subtypes, which we refer to as hPAP alveolar macrophages (hPAP AMs) here. To

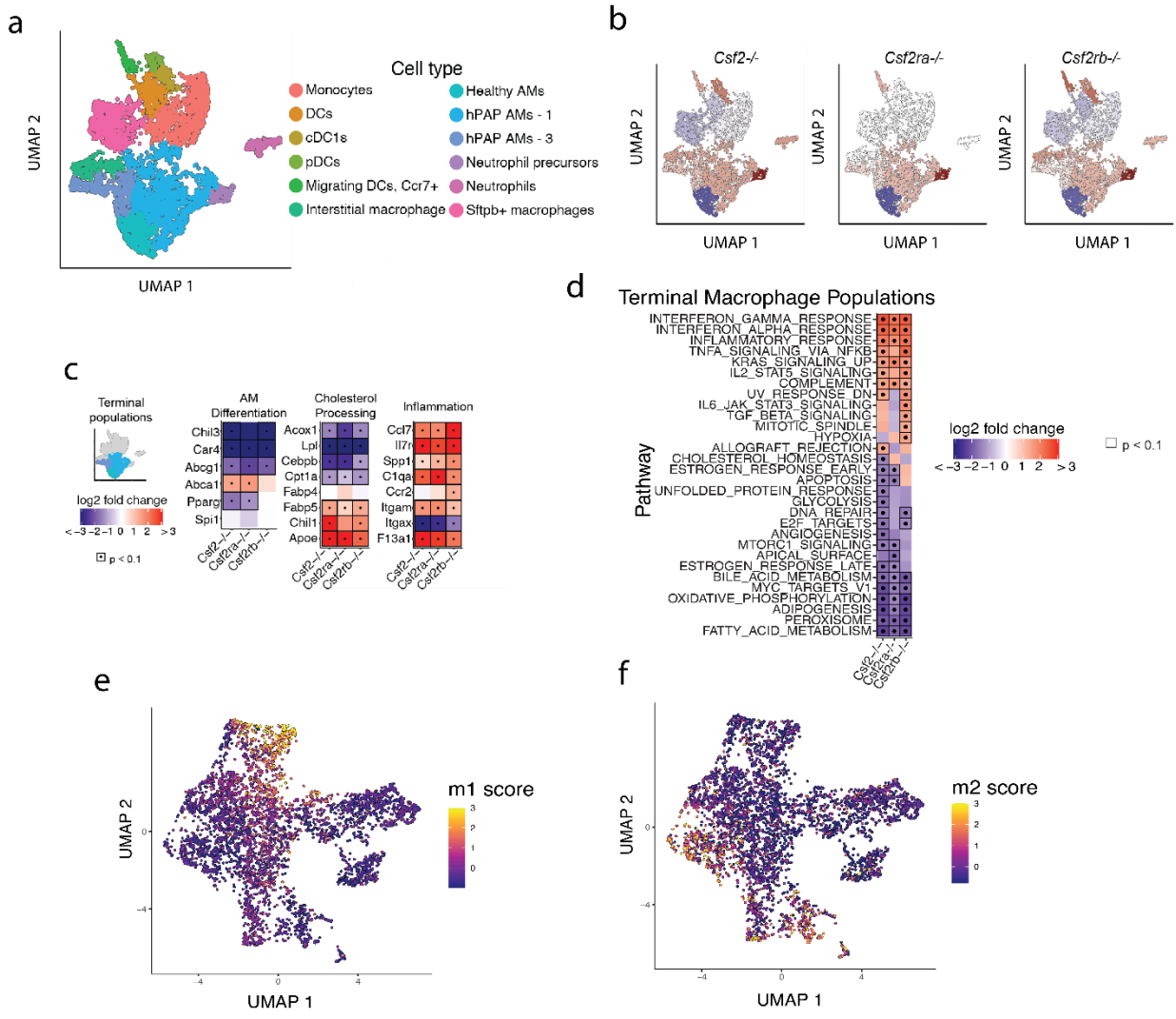

Figure 4-5: **Alterations in macrophage subtypes**. A) A UMAP generated from macrophages across all genotypes, colored by macrophage subtype. B) Enrichment (red) and depletion (blue) for macrophage subtypes in hPAP genotypes compared to wild type. C) Differential expression of genes related to AM differentiation, cholesterol metabolism, and inflammation in terminal macrophage subpopulations. Dots indicates significant with FDR < .1. D) Enrichment/depletion of biological pathways within terminal macrophage populations by genotype, with respect to wild type. E) UMAP of terminal macrophage populations colored by “M1” score or F) “M2” score. See methods.

begin to characterize these hPAP AMs, we first used differential expression to compare the terminal populations in each genotype (healthy AMs in wild type backgrounds and hPAP AMs in

the mutant backgrounds) and identified 465 genes that were mis-regulated in at least one genetic background. Genes that are critical for AM differentiation and surfactant processing were expressed at far lower levels in hPAP AMs than in healthy AMs (Figure 4.5C). An alternative set of cholesterol processing genes including Apoe and Chil1 were upregulated, reminiscent of the profile described for foamy macrophages (Guerrini and Gennaro 2019). Furthermore, hPAP AM populations showed elevated interferon signaling, complement pathway activation, NF-kappa B signaling activity, and antigen presentation signatures relative to healthy AMs (Figure 4.5D). Of the two disease-enriched subtypes, one expressed a strong pro-inflammatory M1-like macrophage signature and the other expressed a strong anti-inflammatory M2-like signature, similar to what we observed in healthy AMs (Figure 4.5E,F) (Jablonski et al. 2015).

The standing model for macrophage dysfunction in PAP is that macrophages are present but immature (B. C. Trapnell et al. 2019), which agrees with our finding of reduced expression of differentiation factors. In healthy individuals the lung is seeded with macrophages during embryonic development and then maintained through local proliferation. Following depletion in injury or viral infection, monocytes are able to extravasate from circulation, infiltrate the lung, and differentiate into alveolar macrophages when that niche is unoccupied (Evren, Ringqvist, and Willinger 2020). Therefore, to assess the differentiation progress of the hPAP AMs, we further subset monocytes and all macrophage populations from our myeloid cells, again projected them into UMAP space, and performed pseudo time trajectory analysis (Figure 4.6A,B). Setting the monocytes as the root, we observe a differentiation trajectory that proceeds from monocytes to healthy alveolar macrophages, with two branches projecting out into two major clusters of hPAP AMs. This suggests that the diseased macrophages occupy an aberrant cell state, rather than simply being immature.

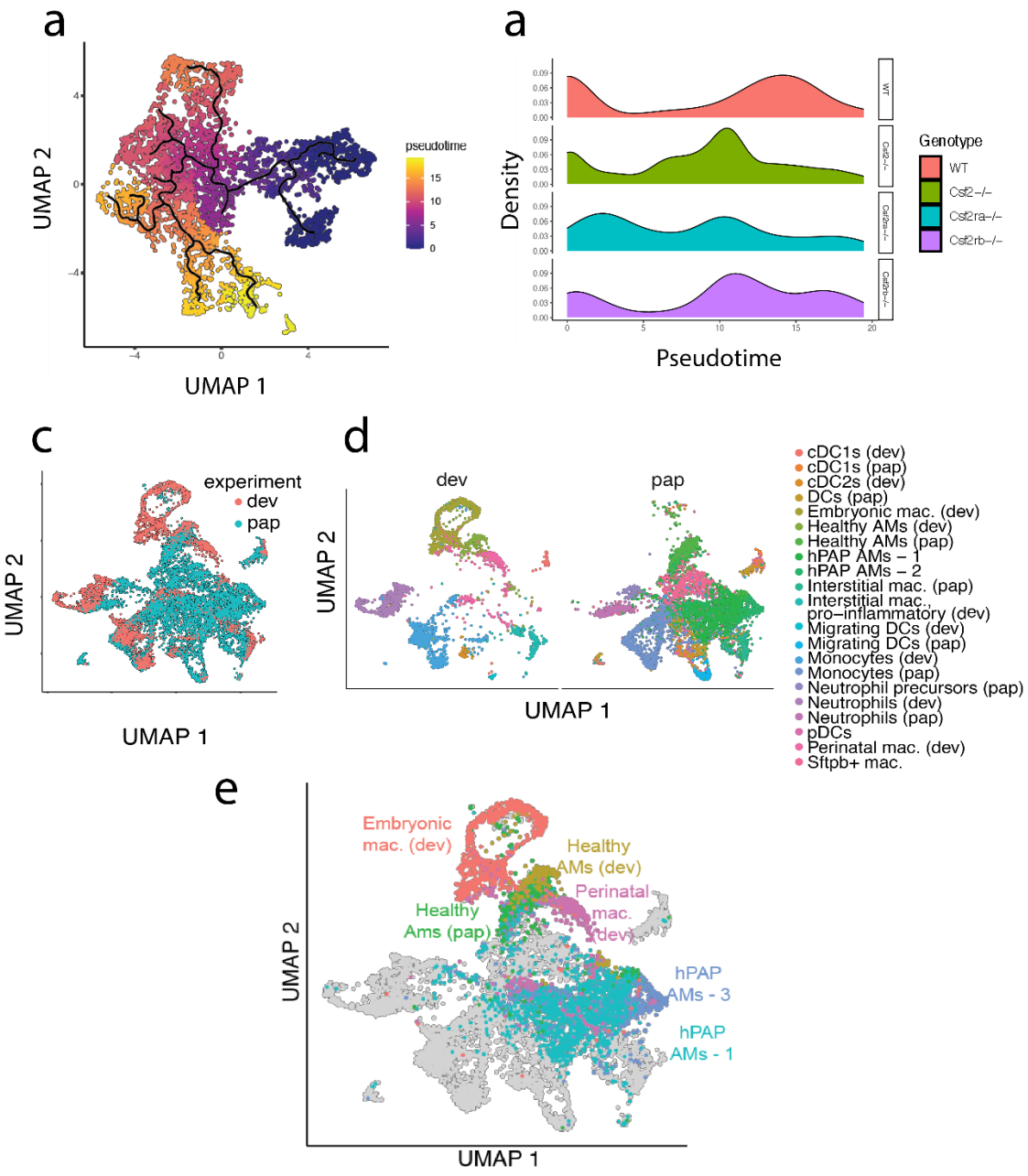

Figure 4-6: **Pseudotime analysis and developmental comparison of macrophage subtypes**. A) Pseudotime trajectory calculated for monocytes and macrophages across all genotypes. B) Distribution of cells by genotype across pseudotime. C) A UMAP from co-embedding all genotypes with a developing mouse lung atlas (see methods). D) UMAP of co-embedding all PAP genotypes and wild type with developmental mouse atlas data, colored by cell subtype and faceted by data source. E) Co-embedding of PAP datasets with developmental mouse atlas data, coloring PAP data only by cell type.

One caveat to this interpretation is that the cells in this experiment were all derived from 12 - 13-week-old mice, so we were unable to directly observe healthy macrophages as they matured in normal development. However, other groups have profiled the murine lung in early development with single cell RNA-seq, and we aligned our myeloid cells to a dataset derived from embryonic (E18.5), perinatal (P1), and mature (P7) mouse lungs (Domingo-Gonzalez et al. 2020) using Seurat. We observe good concordance between defined non-macrophage cell types, including neutrophils, monocytes, and DCs, confirming that the alignment was successful (Figure 4.6C,D). When focusing on alveolar macrophages, we observe similar localization of healthy AMs across the two datasets but very little overlap between hPAP AMs and either perinatal or embryonic AMs from the developmental dataset (Figure 4.6E), further supporting the finding that the hPAP AMs are not occupying a normal developmental state.

The following section is adapted with minimal modifications from a manuscript under preparation. Writing and analysis presented here was led by Dr. Jennifer Franks.

4.3.3 *Silicosis results in alteration sin distinct macrophage subsets and altered cell proportions*

We performed a genomic longitudinal analysis of silicosis in a mouse model using single-nucleus RNA-sequencing from pre- and post- silica instilled mouse lungs. We recovered a total of 23,794 single cells from 12 whole lung samples across four timepoints (Day 0, 7, 28, 56). Thirty-five unique cell states spanning epithelial, endothelial, stromal, myeloid, and lymphoid cell lineages were identified using highly and specifically expressed marker genes (Figure 4.8A, B). Many cell states are differentially abundant over time with most notable changes happening at Day 7 post-exposure (Figure 4.8C). AT1, bronchioalveolar stem cells, regulatory T cells, and

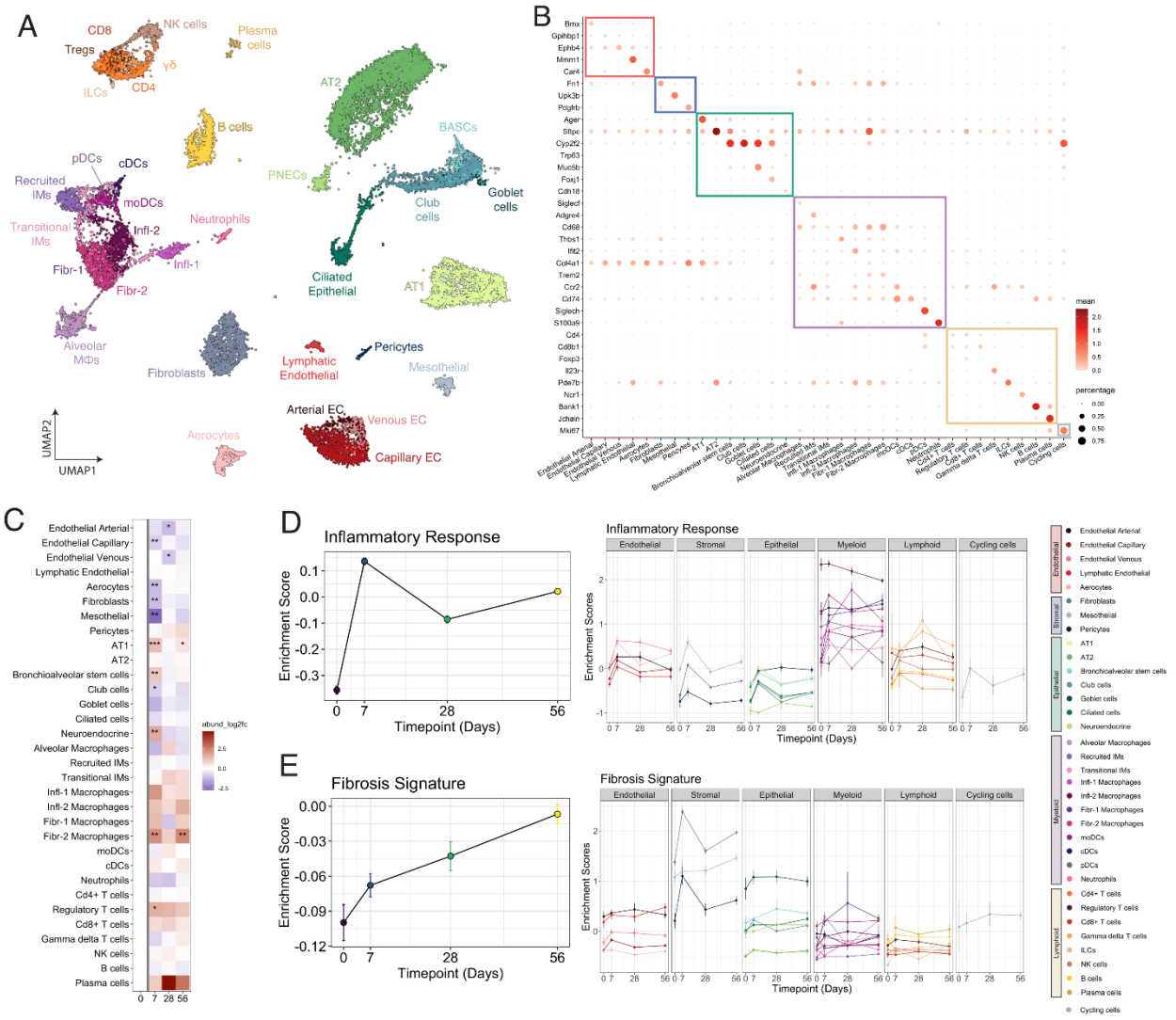

Figure 4-7: **Single-nucleus sequencing of the murine lung pre- and post-intratracheal administration of silica.** (A) 2D UMAP representation of all annotated cell types identified. Each point represents a single cell and is colored according to fine scale annotation. (B) Selected marker genes used for broad and fine-scale cell type annotation. (C) Cell state abundance relative to Day 0 shown over time. *p<0.05, **p<0.005, ***p<0.0005; BH correction. Gene set enrichment scores for Inflammation (D) and Fibrosis (E) shown over time summarized for all cells collected from each timepoint (left panel) and colored by individual cell state (right panel).

Fibr-2 macrophages significantly increase in relative abundance following silica exposure while arterial, capillary, and venous endothelial cells, aerocytes, fibroblasts, mesothelial, and club cells decrease in relative abundance (adjusted p<0.05, beta-binomial test). Neuroendocrine cells demonstrate a significant yet transient increase at Day 7 post-silica exposure which may reflect a

potential role in supporting acute inflammation. Silicosis is known to produce both acute and chronic inflammation in the lung followed by progressive fibrosis. Indeed, gene set enrichment analysis reveals that inflammatory gene processes peak at Day 7 post-exposure and chronically remain higher than baseline while fibrotic gene expression processes continually increase following exposure (Figure 4.8D, E). Cell states uniquely vary both in the baseline activation of inflammatory and fibrotic gene signatures and in the subsequent temporal response following silica exposure.

We further interrogated the myeloid cells to identify the genetic programs activated in this lineage during silica exposure in the mouse lung (Figure 4.8A). Within the myeloid cells, we identified previously described populations of alveolar (SiglecF+) and interstitial (SiglecF-) macrophages as well as populations of dendritic cells and neutrophils (Figure 4.8B). Interstitial macrophages were highly abundant in the silicosis lung and demonstrated remarkable heterogeneity. Recruited IMs represent the most recently recruited cells in the lung based on high expression of CCR2. Transitional IMs express low levels of many pro-inflammatory and pro-fibrotic genes and likely represent an intermediate cell state poised to adopt a more specialized macrophage subtype phenotype. All macrophages in our analysis show some activation of both inflammatory and fibrotic gene processes, and the traditional M1/M2 nomenclature is insufficient to describe the heterogeneity evident in our dataset (Figure 4.8C-E). Thus, we annotated two populations of pro-inflammatory macrophages (Infl-1, Infl-2) and two populations of pro-fibrotic macrophages (Fibr-1, Fibr-2) based on gene expression and pathway activation differences. Infl-1 showed higher activation of TNF related gene expression and Infl-2 shows gene expression more skewed toward IFNγ. Infl-1 peaks at Day 7 and likely represents a population of myeloid cells responsible for acute inflammation in the silicosis lung, whereas Infl-

2 macrophages contribute broad inflammatory signatures evident in chronic inflammation. Infl-2 macrophages also express many pro-fibrotic genes and may serve as precursors to other fibrotic macrophage cell states. Fibr-1 macrophages likely drive fibrosis due to high expression of tissue-

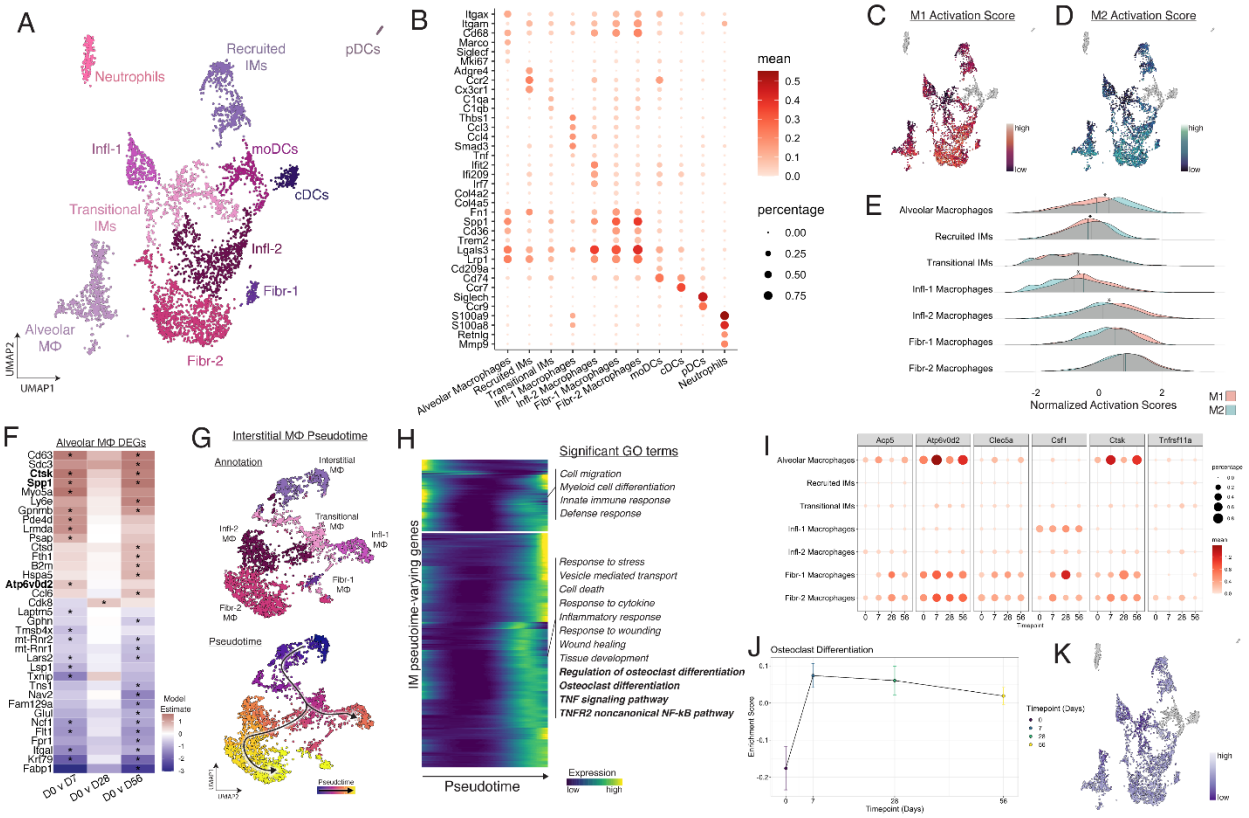

Figure 4-8: **Myeloid cells demonstrate activation of osteoclast related transcriptional programs**. (A) 2D UMAP representation of myeloid cells. Each point represents a single cell and is colored according to fine scale annotation. (B) Selected marker genes used for fine-scale cell state annotation of myeloid cells. (C) M1 and (D) M2 gene sets were used to calculate an activation score for each cell. Each point is a cell colored according to the relative intensity of the signature. (E) M1 and M2 activation scores for each of the fine-scale cell state annotations was plotted in a histogram and tested for differences in mean (paired T-test). *p<0.05; BH correction. (F) Differentially expressed genes relative to Day 0 were identified for the alveolar macrophages. Genes in bold typeface are related to osteoclast differentiation and development. *p<0.05, FDR corrected. (G) 2D UMAP of interstitial macrophages colored according to fine-scale annotation (top panel) and pseudotime (bottom panel). (F) Heatmap of genes differentially expressed over pseudotime as interstitial macrophages differentiate toward pro-fibrotic phenotypes. Each row represents a gene and color represents relative intensity of expression. Select significant (q<0.05, gprofiler) representative gene ontology terms for each cluster of genes are shown on the right. (I) Expression of osteoclast genes plotted over time for each of the macrophage subsets. (J) Gene set enrichment scores for Osteoclast Differentiation were calculated for all macrophages and plotted over time. (K) Macrophages colored according to relative intensity of Osteoclast Differentiation enrichment score.

remodeling genes including Col4a1, Col4a2, and Fn1. The Fibr-2 macrophage subset likely represents end-stage pro-fibrotic macrophages in the silicosis lung due to waning expression of pro-fibrotic genes coupled with increased expression of foamy macrophage markers including Spp1, Cd36, and lipid-metabolizing genes such as Lgals3 and Lrp1. Many pro-fibrotic genes including Spp1, Lgals3, and Lrp1 are also highly expressed in alveolar macrophages.

To identify the temporal changes in genetic programs associated with silicosis, we performed two analyses. First, in the SiglecF+ alveolar macrophages we identified differentially expressed genes over time. Several osteoclast marker genes were significantly increased in the alveolar macrophages of silicosis lungs compared to baseline, including Ctsk, Spp1, and Atp6v0d2 (Figure 4.8F). To identify the genetic programs changing within the interstitial macrophages (SiglecF-), we performed pseudotime analysis (Figure 4.8G). Pseudotime analysis links the progression of recently recruited interstitial macrophages to a transitional state followed by pro-inflammatory and pro-fibrotic phenotypes. Genes with significantly varying expression along the pseudotime trajectory were identified, hierarchically clustered, and clusters were annotated based on gene ontology (Figure 4.8H). Genes that show increased expression later in pseudotime are significantly enriched for inflammatory response, wound healing, tissue development, osteoclast differentiation, and TNF signaling pathway (adjusted $p<0.05$, gProfiler). We calculated an enrichment score for osteoclast differentiation for each macrophage and plotted this over time (Fig 1J-K). We can see the activation of this signature increases dramatically post-silica exposure and the highest enrichment scores are present in alveolar macrophages and pro-fibrotic macrophage populations Fibr-1 and Fibr-2. Moreover, these populations demonstrate robust expression of osteoclast markers including Acp5, Atp6v062, Clec5a, Ctsk and moderate expression of Csf1 and Tnfrsf11a (Fig 4I).

In summary, both tissue-resident alveolar macrophages and infiltrating bone-marrow derived macrophages demonstrate robust genomic signatures of osteoclastic transformation following silica exposure. This supports that microenvironmental signals are critical in shaping pulmonary macrophage polarization states and functional phenotype.

4.4 DISCUSSION

In analysis of PAP we characterized alterations of macrophage populations in the lung. Given the role of GM-CSF signaling in alveolar macrophage specification and maintenance (Shibata et al. 2001), loss of alveolar macrophages in diseased lungs (Figure 4.4) was expected. However, it was not obvious if we should expect purely an accumulation of alveolar macrophage precursors, loss of alveolar macrophages due to cell death, or diversion to an alternate phenotype distinct from mature alveolar macrophages. Our findings are consistent with a model in which GM-CSF signaling disruption causes an aberrant differentiation pathway in which would-be alveolar macrophages fail to differentiate but survive in an altered state defined by expression of transcripts in cholesterol processing and inflammatory pathways (Figure 4.5C). Beyond better understanding an aspect of macrophage development, this observation may be clinically relevant in the context of proposed “macrophage transplantation therapy” (Takuji Suzuki et al. 2014) aiming to transfer genetically engineered macrophages into PAP patients in which transplanted macrophages are intended to fill the niche of alveolar macrophages. The presence of improperly differentiated macrophages (Figure 4.6) in PAP lungs suggests that the success of transplantation therapies will depend not only on the ability of added cells to fill the roles played by non-existent alveolar macrophages, but in the ability to overcome any pathological effects of aberrantly differentiated macrophages already present in the lung.

In the silicosis model, we find shifts consistent with emergence of osteoclast-like macrophages in the diseased lung (Figure 4.8) in addition to activation states featuring inflammatory and pro-fibrotic gene programs (Figure 4.8B). A spectrum of pro-inflammatory and pro-fibrotic responses is a hallmark of diverse fibrotic diseases, so characterization of the specific alterations seen in this disease model (Figure 4.7 D,E; Figure 4.8A,B) further the understanding of which programs – at which time points - are activated in silica exposure particularly. More surprisingly, the detection of an osteoclast-like signature in lung macrophages – apparently both from alveolar- and interstitial-derived populations (Figure 4.8I) – represents a striking and potentially impactful aspect of disease. Adoption of an osteoclast-like gene program by lung macrophages has been recently observed in a distinct lung particulate disease model (Uehara et al. 2021), while our data suggests that macrophages take on a phenotype typically adopted to drive bone degradation in response to mineral particulates in the lung. This would represent a fascinating example of plasticity in macrophages (Evren, Ringqvist, and Willinger 2020) with adoption of a bone-resident macrophage phenotype within the lung. Therapeutically, work is already underway to understand if disruption of osteoclast-like macrophages via blockage of RANK-L – a key mediator of osteoclast specification (Park, Lee, and Lee 2017) – could reduce disease severity in silicosis and related ailments via blocking osteoclast-like phenotype adoption and resulting release of catabolic compounds.

In mouse models of PAP and silicosis we find evidence of widespread alterations in cell type abundances. While our analysis has initially focused on alterations in macrophage subtypes – due to the known effects of GM-CSF deficiency on macrophages in PAP (B. C. Trapnell et al. 2019) and the essential role of macrophages for debris clearance in the lung (Thakur et al. 2009; Huaux 2007; Hamilton, Thakur, and Holian 2008) in silicosis - alterations in additional cell

types suggests that significant behavior occurs in a variety of cell types. For example, future work can interrogate the extent to which we see evidence of AT2-to-AT1 differentiation through a particular Krt8 intermediate that mirrors alterations seen in a bleomycin murine model of lung injury (Strunz et al. 2020). Similarly, it will be illuminating to see the extent to which alterations in immune cells represent a non-specific inflammatory response to many injuries, or a specific alteration that varies by the type of insult.

Chapter 5. CONCLUSIONS

The work in this dissertation falls largely under one of two distinct umbrellas: the use of machine learning models to understand gene regulation (Chapter 2 and Chapter 3.3.5) or the optimization and application of single-cell methods to tissue physiology. (Chapter 3 and 4).

5.1.1 *Predictive models of gene expression*

In the context of *P. falciparum*, model inspection suggests surprising findings like a low predictive value for AP2 TF motifs (Figure 2.2C), low utility of TSS-centric epigenomic features (Figure 2.3B), and divergent utility of covalent histone modifications between different bloodborne stages (Figure 2.4A). Differences in use of histone modifications is particularly striking, given that in similar modeling of mammalian expression an individual histone mark would be of generally consistent utility in models across cell types (Cheng et al. 2011; Dong et al. 2012). If the finding of stage-specific histone modification importance is validated in the future, it could suggest that stage-specific activity of chromatin readers occurs at a genome-wide scale in contrast to paradigms of expression regulation that have historically focused on context-specific activity of sequence-specific TFs driving cell type identity and cellular responses (Bhattacharjee et al. 2013; Mullen et al. 2011; Flitsch, Laupman, and Brüstle 2020; M. Xu et al. 2022; Lee et al. 2012; Hosokawa and Rothenberg 2021).

In the cells of the human heart, fitting a linear model of RNA expression reaffirms the informative value of distal DNA sequences across primary cell types (Figure 3.5C), quantifies the extent to which a simple TF motif presence/absence code explains cell-type-specific expression (Figure 3.5C), and quantifies the relationship between motifs and expression (Figure 3.5G). Quantifying these relationships is important given past and ongoing work in the field of gene

regulation to understand cell type identity and response to stimuli in terms of the activity of particular TFs. For example, recent human body-wide study of chromatin accessibility has analyzed cell identity in relation to enrichment or depletion of TF motifs in accessible chromatin (Domcke et al. 2020; K. Zhang et al. 2021). As the field advances, it is natural to extend statements like "TF X promotes expression in context Y", towards a full accounting of variation like "TF X drives a 2-fold increase in expression in context Y". Even fitting simple models – such as a linear model using binary motif features – moves towards this goal, advancing beyond tests of TF alterations compared to a null model of no variation.

Future extensions of work such as our analysis of *Plasmodium* or human heart gene expression will almost certainly entail use of deep learning methods. Deep neural networks have generated state-of-the-art performance in gene expression prediction for several years (Avsec et al. 2021; Agarwal and Shendure 2020; J. Zhou et al. 2018; Y. Chen et al. 2016; Singh et al. 2016), suggesting they are able to incorporate subtle features or interactions that shallow learning methods have been unable to utilize. As inspection of a model's feature use is only as informative as the model's ability to predict transcription, future work of more highly accurate models represents a clear future direction. Particularly as expression models become ever more accurate, it will be exciting to see the extent to which the grammar of expression regulation can eventually be simplified or understood in terms of simple rules, or if expression control is inherently a highly complex process. For instance, significant work in biology understands transcription in terms of a small set of key regulators affecting changes in transcription (Bhattacharjee et al. 2013), an assumption well matched by models such as the work in Chapter 2 and 3.3.5 where relatively simple features are related to expression. However, it is possible that the true grammar of transcription is not faithfully distilled into a set of readily enumerated rules

but rather a complex process driven by numerous small contributions – akin to how structure in naturally occurring proteins are typically an emergent property of many pair-wise interactions between residues, rather than locked in by a small handful of elements (Dill and MacCallum 2012). Once the field develops a model that can effectively quantify expression variation, we will be in a much stronger position to understand the extent to which a highly accurate but still simple model of expression is possible, or purely an aspiration of biologists wanting to simplify a process that inherently cannot be so easily conveyed.

5.1.2 *Development of single cell methods*

At the point at which the work in this dissertation began, a significant challenge existed in applying sci RNA-Seq to solid mammalian tissues, a context where the advantages of combinatorial indexing were clear. Use of split-pool methods for tissue biology was exciting due to aspects like profiling many cells to get coverage of low-abundance cell types, obtaining statistical power by getting many pseudo replicates in the form of many cells per tissue, and executing more elaborate experimental designs in terms of unique tissues profiled for a feasible cost. However, initial attempts to directly apply previous sci RNA-Seq nuclei preparation methods (Cao et al. 2019, 2017) were unsuccessful in solid mouse tissues outside of the brain. Work described in this dissertation in chapter 4 (Figure 4.1) represented a successful "Version 1.0" of running sci RNA-Seq preparations in mammalian tissue, facilitating an early tranche of studies (the human heart and two mouse lung analyses described in this dissertation). However, the method suffered from notable limitations, including the need for a FACS sorting step that precluded massive throughput experiments and an inability to obtain high quality data from all human tissue contexts. In parallel, methods developed in another group at UW represent a newer, improved workflow that does not require FACS sorting and has been applied to a wider array of

tissue contexts (B. K. Martin et al. 2021). While that method appears to be the basis of near future work and sci RNA-Seq application, our previous optimization work both facilitated the generation of an initial series of datasets and provided insight into areas that are key for successful nuclei preparation.

5.1.3 *Studying tissue physiology and pathology using snRNA-Seq*

In human hearts, our work suggests considerable correlation between age or sex and expression of numerous pathways, chromatin accessibility, and cell type composition of tissue (Chapter 3). These changes include evidence of alterations by sex in TGFβ signaling and metabolic rewiring, as well as variation by age in inflammatory and immune-activation pathways. We see indications of these changes both in transcriptional (Figure 3.2 and 3.3) and ATAC-Seq (Figure 3.2F; Figure 3.3C) data. As noted in Chapter 3, this analysis lacks generalizability due to the limited number of donors profiled, making it impossible to rule out confounding effects from unmeasured traits like donor exercise levels, latent comorbidities, and so on. Ongoing and future work aims to extend these analysis strategies to larger cohorts.

Identifying molecular hallmarks of age- or sex-dependent variation would be of keen interest for therapeutic development. For instance, age is the single greatest risk factor for heart disease (Rodgers et al. 2019) so understanding the pathways that correlate with age represents a crucial first step in eventually identifying causal aspects of pathological aging. Our analysis strategy inherently finds these correlations in a cell-type resolved manner which lets future work focus on variation in cell types relevant to a disease of interest. For example, defining age-dependent variation in fibroblasts will be of greater interest to work on cardiac fibrosis (Aghajanian et al. 2019) whereas variation in vascular endothelial cells would be of more targeted value to those studying coronary artery disease (Howe and Fish 2019). Our preliminary

findings of pathways altered by age or sex suggest that many such areas of variation will be clinically targetable given already underway research into cardiac therapies via metabolic modulation (Lopatin 2015), targeting of inflammation (Murphy et al. 2020), and inhibiting TGFβ-driven fibrosis (Meng, Nikolic-Paterson, and Lan 2016).

In two models of mouse lung disease, we identify alterations in cell type proportions and expression programs. In PAP, we identify molecular signatures of macrophages that have failed to properly differentiate due to disrupted GM-CSF signaling (Figure 4.6). In silicosis, we identify subsets of macrophages occurring in the diseased lung (Figure 4.8), alterations in numerous other cell types (Figure 4.7C), and find evidence of an osteoclast-like gene expression program in macrophages in the lung (Figure 4.8F,I,J). In the context of these particular diseases we extend our understanding of alterations at cellular resolution and transcriptome-wide scale and help guide further experiments, such as now-underway work to study alterations in B-cell proportions in the PAP lung and evaluate the effects of osteoclast-differentiation-blocking agents in inhalation lung injury models. More broadly, these datasets and analyses feed into broader efforts to understand signatures of various diseases at single-cell resolution. Similarly to efforts building databases of perturbation signatures using bulk RNA-Seq (Lamb et al. 2006), characterizing the effects of perturbations in animal systems via single-cell data will reveal signatures in terms of cell type proportion changes or differential expression in particular cell types. Those signatures will be of great value for understanding the extent to which animal models of disease do or do not match alterations in corresponding human disease, the correspondence between diseases of unknown etiology and datasets of known perturbations, and the degree of overlap in alterations between injury models.

BIBLIOGRAPHY

Abbate, Antonio, Stefano Toldo, Carlo Marchetti, Jordana Kron, Benjamin W. Van Tassell, and Charles A. Dinarello. 2020. "Interleukin-1 and the Inflammasome as Therapeutic Targets in Cardiovascular Disease." *Circulation Research* 126 (9): 1260–80.

Adams, Taylor S., Jonas C. Schupp, Sergio Poli, Ehab A. Ayaub, Nir Neumark, Farida Ahangari, Sarah G. Chu, et al. 2020. "Single-Cell RNA-Seq Reveals Ectopic and Aberrant Lung-Resident Cell Populations in Idiopathic Pulmonary Fibrosis." *Science Advances* 6 (28): eaba1983.

Adjalley, Sophie H., Christophe D. Chabbert, Bernd Klaus, Vicent Pelechano, and Lars M. Steinmetz. 2016. "Landscape and Dynamics of Transcription Initiation in the Malaria Parasite Plasmodium Falciparum." *Cell Reports* 14 (10): 2463–75.

Agarwal, Vikram, and Jay Shendure. 2020. "Predicting MRNA Abundance Directly from Genomic Sequence Using Deep Convolutional Neural Networks." *Cell Reports* 31 (7): 107663.

Aghajanian, Haig, Toru Kimura, Joel G. Rurik, Aidan S. Hancock, Michael S. Leibowitz, Li Li, John Scholler, et al. 2019. "Targeting Cardiac Fibrosis with Engineered T Cells." *Nature* 573 (7774): 430–33.

Akerberg, Brynn N., Fei Gu, Nathan J. VanDusen, Xiaoran Zhang, Rui Dong, Kai Li, Bing Zhang, et al. 2019. "A Reference Map of Murine Cardiac Transcription Factor Chromatin Occupancy Identifies Dynamic and Conserved Enhancers." *Nature Communications* 10 (1): 4907.

Alexanian, Michael, Pawel F. Przytycki, Rudi Micheletti, Arun Padmanabhan, Lin Ye, Joshua G. Travers, Barbara Gonzalez-Teran, et al. 2021. "A Transcriptional Switch Governs Fibroblast Activation in Heart Disease." *Nature* 595 (7867): 438–43.

Angelidis, Ilias, Lukas M. Simon, Isis E. Fernandez, Maximilian Strunz, Christoph H. Mayr, Flavia R. Greiffo, George Tsitsiridis, et al. 2019. "An Atlas of the Aging Lung Mapped by Single Cell Transcriptomics and Deep Tissue Proteomics." *Nature Communications* 10 (1): 963.

Aran, Dvir, Zicheng Hu, and Atul J. Butte. 2017. "XCell: Digitally Portraying the Tissue Cellular Heterogeneity Landscape." *Genome Biology* 18 (1): 220.

Archin, Nancie M., Jennifer L. Kirchherr, Julia Am Sung, Genevieve Clutton, Katherine Sholtis, Yinyan Xu, Brigitte Allard, et al. 2017. "Interval Dosing with the HDAC Inhibitor Vorinostat Effectively Reverses HIV Latency." *The Journal of Clinical Investigation* 127 (8): 3126–35.

Aurrecoechea, Cristina, John Brestelli, Brian P. Brunk, Jennifer Dommer, Steve Fischer, Bindu Gajria, Xin Gao, et al. 2009. "PlasmoDB: A Functional Genomic Database for Malaria Parasites." *Nucleic Acids Research* 37 (Database issue): D539-43.

Avsec, Žiga, Vikram Agarwal, Daniel Visentin, Joseph R. Ledsam, Agnieszka Grabska-Barwinska, Kyle R. Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R. Kelley. 2021. "Effective Gene Expression Prediction from Sequence by Integrating Long-Range Interactions." *Nature Methods* 18 (10): 1196–1203.

Ay, Ferhat, Evelien M. Bunnik, Nelle Varoquaux, Sebastiaan M. Bol, Jacques Prudhomme, Jean-Philippe Vert, William Stafford Noble, and Karine G. Le Roch. 2014. "Three-Dimensional Modeling of the P. Falciparum Genome during the Erythrocytic Cycle

Reveals a Strong Connection between Genome Architecture and Gene Expression." *Genome Research* 24 (6): 974–88.

Ay, Ferhat, Evelien M. Bunnik, Nelle Varoquaux, Jean-Philippe Vert, William Stafford Noble, and Karine G. Le Roch. 2015. "Multiple Dimensions of Epigenetic Gene Regulation in the Malaria Parasite Plasmodium Falciparum: Gene Regulation via Histone Modifications, Nucleosome Positioning and Nuclear Architecture in P. Falciparum." *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 37 (2): 182–94.

Bailey, Timothy L., Mikael Boden, Fabian A. Buske, Martin Frith, Charles E. Grant, Luca Clementi, Jingyuan Ren, Wilfred W. Li, and William S. Noble. 2009. "MEME SUITE: Tools for Motif Discovery and Searching." *Nucleic Acids Research* 37 (Web Server issue): W202-8.

Balaji, S., M. Madan Babu, Lakshminarayan M. Iyer, and L. Aravind. 2005. "Discovery of the Principal Specific Transcription Factors of Apicomplexa and Their Implication for the Evolution of the AP2-Integrase DNA Binding Domains." *Nucleic Acids Research* 33 (13): 3994–4006.

Barber, Rina Foygel, and Emmanuel J. Candès. 2015. "Controlling the False Discovery Rate via Knockoffs." *Annals of Statistics* 43 (5). https://doi.org/10.1214/15-aos1337.

Barnes, Hayley, Nicole S. L. Goh, Tracy L. Leong, and Ryan Hoy. 2019. "Silica-Associated Lung Disease: An Old-World Exposure in Modern Industries." *Respirology (Carlton, Vic.)* 24 (12): 1165–75.

Bártfai, Richárd, Wieteke A. M. Hoeijmakers, Adriana M. Salcedo-Amaya, Arne H. Smits, Eva Janssen-Megens, Anita Kaan, Moritz Treeck, Tim-Wolf Gilberger, Kees-Jan Françoijs, and Hendrik G. Stunnenberg. 2010. "H2A.Z Demarcates Intergenic Regions of the Plasmodium Falciparum Epigenome That Are Dynamically Marked by H3K9ac and H3K4me3." *PLoS Pathogens* 6 (12): e1001223.

Batlle, Eduard, and Joan Massagué. 2019. "Transforming Growth Factor-β Signaling in Immunity and Cancer." *Immunity* 50 (4): 924–40.

Beale, Anna L., Philippe Meyer, Thomas H. Marwick, Carolyn S. P. Lam, and David M. Kaye. 2018. "Sex Differences in Cardiovascular Pathophysiology: Why Women Are Overrepresented in Heart Failure With Preserved Ejection Fraction." *Circulation* 138 (2): 198–205.

Bedogni, Francesco, Rebecca D. Hodge, Gina E. Elsen, Branden R. Nelson, Ray A. M. Daza, Richard P. Beyer, Theo K. Bammler, John L. R. Rubenstein, and Robert F. Hevner. 2010. "Tbr1 Regulates Regional and Laminar Identity of Postmitotic Neurons in Developing Neocortex." *Proceedings of the National Academy of Sciences of the United States of America* 107 (29): 13129–34.

Beer, Michael A., and Saeed Tavazoie. 2004. "Predicting Gene Expression from Sequence." *Cell* 117 (2): 185–98.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 57 (1): 289–300.

Bhattacharjee, Soumya, Kaushik Renganaath, Rajesh Mehrotra, and Sandhya Mehrotra. 2013. "Combinatorial Control of Gene Expression." *BioMed Research International* 2013 (August): 407263.

Blatti, Charles, Majid Kazemian, Scot Wolfe, Michael Brodsky, and Saurabh Sinha. 2015. "Integrating Motif, DNA Accessibility and Gene Expression Data to Build Regulatory Maps in an Organism." *Nucleic Acids Research* 43 (8): 3998–4012.

Boer, Carl G. de, and Timothy R. Hughes. 2012. "YeTFaSCo: A Database of Evaluated Yeast Transcription Factor Sequence Specificities." *Nucleic Acids Research* 40 (Database issue): D169-79.

Bonasio, Roberto. 2015. "The Expanding Epigenetic Landscape of Non-Model Organisms." *The Journal of Experimental Biology* 218 (Pt 1): 114–22.

Borie, R., C. Danel, M-P Debray, C. Taille, M-C Dombret, M. Aubier, R. Epaud, and B. Crestani. 2011. "Pulmonary Alveolar Proteinosis." *European Respiratory Review: An Official Journal of the European Respiratory Society* 20 (120): 98–107.

Bunnik, Evelien M., Kate B. Cook, Nelle Varoquaux, Gayani Batugedara, Jacques Prudhomme, Anthony Cort, Lirong Shi, et al. 2018. "Changes in Genome Organization of Parasite-Specific Gene Families during the Plasmodium Transmission Stages." *Nature Communications* 9 (1): 1–15.

Bunnik, Evelien M., Anton Polishko, Jacques Prudhomme, Nadia Ponts, Sarjeet S. Gill, Stefano Lonardi, and Karine G. Le Roch. 2014. "DNA-Encoded Nucleosome Occupancy Is Associated with Transcription Levels in the Human Malaria Parasite Plasmodium Falciparum." *BMC Genomics* 15 (1): 347.

Bussemaker, H. J., H. Li, and E. D. Siggia. 2001. "Regulatory Element Detection Using Correlation with Expression." *Nature Genetics* 27 (2): 167–71.

Campbell, Tracey L., Erandi K. De Silva, Kellen L. Olszewski, Olivier Elemento, and Manuel Llinás. 2010. "Identification and Genome-Wide Prediction of DNA Binding Specificities for the ApiAP2 Family of Regulators from the Malaria Parasite." *PLoS Pathogens* 6 (10): e1001165.

Cao, Junyue, Jonathan S. Packer, Vijay Ramani, Darren A. Cusanovich, Chau Huynh, Riza Daza, Xiaojie Qiu, et al. 2017. "Comprehensive Single-Cell Transcriptional Profiling of a Multicellular Organism." *Science (New York, N.Y.)* 357 (6352): 661–67.

Cao, Junyue, Malte Spielmann, Xiaojie Qiu, Xingfan Huang, Daniel M. Ibrahim, Andrew J. Hill, Fan Zhang, et al. 2019. "The Single-Cell Transcriptional Landscape of Mammalian Organogenesis." *Nature* 566 (7745): 496–502.

Capogrossi, Maurizio C. 2004. "Cardiac Stem Cells Fail with Aging: A New Mechanism for the Age-Dependent Decline in Cardiac Function." *Circulation Research*.

Casamassimi, Amelia, and Alfredo Ciccodicola. 2019. "Transcriptional Regulation: Molecules, Involved Mechanisms, and Misregulation." *International Journal of Molecular Sciences* 20 (6): 1281.

Cassel, Suzanne L., Stephanie C. Eisenbarth, Shankar S. Iyer, Jeffrey J. Sadler, Oscar R. Colegio, Linda A. Tephly, A. Brent Carter, Paul B. Rothman, Richard A. Flavell, and Fayyaz S. Sutterwala. 2008. "The Nalp3 Inflammasome Is Essential for the Development of Silicosis." *Proceedings of the National Academy of Sciences of the United States of America* 105 (26): 9035–40.

Chaffin, Mark, Irinna Papangeli, Bridget Simonson, Amer-Denis Akkad, Matthew C. Hill, Alessandro Arduini, Stephen J. Fleming, et al. 2022. "Single-Nucleus Profiling of Human Dilated and Hypertrophic Cardiomyopathy." *Nature* 608 (7921): 174–80.

Chen, Tianqi, and Carlos Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." *ArXiv [Cs.LG]*. arXiv. http://arxiv.org/abs/1603.02754.

Chen, Yifei, Yi Li, Rajiv Narayan, Aravind Subramanian, and Xiaohui Xie. 2016. “Gene Expression Inference with Deep Learning.” *Bioinformatics (Oxford, England)* 32 (12): 1832–39.

Cheng, Chao, Koon-Kiu Yan, Kevin Y. Yip, Joel Rozowsky, Roger Alexander, Chong Shou, and Mark Gerstein. 2011. “A Statistical Framework for Modeling Gene Expression Using Chromatin Features and Application to ModENCODE Datasets.” *Genome Biology* 12 (2): R15.

Chookajorn, Thanat, Ron Dzikowski, Matthias Frank, Felomena Li, Alisha Z. Jiwani, Daniel L. Hartl, and Kirk W. Deitsch. 2007. “Epigenetic Memory at Malaria Virulence Genes.” *Proceedings of the National Academy of Sciences of the United States of America* 104 (3): 899–902.

Colombier, Pauline, Boris Halgand, Claire Chédeville, Caroline Chariau, Valentin François-Campion, Stéphanie Kilens, Nicolas Vedrenne, et al. 2020. “NOTO Transcription Factor Directs Human Induced Pluripotent Stem Cell-Derived Mesendoderm Progenitors to a Notochordal Fate.” *Cells* 9 (2). https://doi.org/10.3390/cells9020509.

Conlon, Erin M., X. Shirley Liu, Jason D. Lieb, and Jun S. Liu. 2003. “Integrating Regulatory Motif Discovery and Genome-Wide Expression Analysis.” *Proceedings of the National Academy of Sciences of the United States of America* 100 (6): 3339–44.

Conway, Bryan R., Eoin D. O’Sullivan, Carolynn Cairns, James O’Sullivan, Daniel J. Simpson, Angela Salzano, Katie Connor, et al. 2020. “Kidney Single-Cell Atlas Reveals Myeloid Heterogeneity in Progression and Regression of Kidney Disease.” *Journal of the American Society of Nephrology: JASN* 31 (12): 2833–54.

Coppé, Jean-Philippe, Pierre-Yves Desprez, Ana Krtolica, and Judith Campisi. 2010. “The Senescence-Associated Secretory Phenotype: The Dark Side of Tumor Suppression.” *Annual Review of Pathology* 5: 99–118.

Coulson, Richard M. R., Neil Hall, and Christos A. Ouzounis. 2004. “Comparative Genomics of Transcriptional Control in the Human Malaria Parasite Plasmodium Falciparum.” *Genome Research* 14 (8): 1548–54.

Cusanovich, Darren A., Riza Daza, Andrew Adey, Hannah A. Pliner, Lena Christiansen, Kevin L. Gunderson, Frank J. Steemers, Cole Trapnell, and Jay Shendure. 2015. “Multiplex Single Cell Profiling of Chromatin Accessibility by Combinatorial Cellular Indexing.” *Science (New York, N.Y.)* 348 (6237): 910–14.

Cusanovich, Darren A., Andrew J. Hill, Delasa Aghamirzaie, Riza M. Daza, Hannah A. Pliner, Joel B. Berletch, Galina N. Filippova, et al. 2018. “A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility.” *Cell* 174 (5): 1309-1324.e18.

De Silva, Erandi K., Andrew R. Gehrke, Kellen Olszewski, Ilsa León, Jasdave S. Chahal, Martha L. Bulyk, and Manuel Llinás. 2008. “Specific DNA-Binding by Apicomplexan AP2 Transcription Factors.” *Proceedings of the National Academy of Sciences of the United States of America* 105 (24): 8393–98.

De Val, Sarah, and Brian L. Black. 2009. “Transcriptional Control of Endothelial Cell Development.” *Developmental Cell* 16 (2): 180–95.

DeLong, E. R., D. M. DeLong, and D. L. Clarke-Pearson. 1988. “Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach.” *Biometrics* 44 (3): 837–45.

Desjardins, Cody A., and Francisco J. Naya. 2016. “The Function of the MEF2 Family of Transcription Factors in Cardiac Development, Cardiogenomics, and Direct

Reprogramming." *Journal of Cardiovascular Development and Disease* 3 (3). https://doi.org/10.3390/jcdd3030026.
Dill, Ken A., and Justin L. MacCallum. 2012. "The Protein-Folding Problem, 50 Years On." *Science (New York, N.Y.)* 338 (6110): 1042–46.
Domcke, Silvia, Andrew J. Hill, Riza M. Daza, Junyue Cao, Diana R. O'Day, Hannah A. Pliner, Kimberly A. Aldinger, et al. 2020. "A Human Cell Atlas of Fetal Chromatin Accessibility." *Science* 370 (6518). https://doi.org/10.1126/science.aba7612.
Domingo-Gonzalez, Racquel, Fabio Zanini, Xibing Che, Min Liu, Robert C. Jones, Michael A. Swift, Stephen R. Quake, David N. Cornfield, and Cristina M. Alvira. 2020. "Diverse Homeostatic and Immunomodulatory Roles of Immune Cells in the Developing Mouse Lung at Single Cell Resolution." *ELife* 9 (June). https://doi.org/10.7554/eLife.56890.
Dong, Xianjun, Melissa C. Greven, Anshul Kundaje, Sarah Djebali, James B. Brown, Chao Cheng, Thomas R. Gingeras, et al. 2012. "Modeling Gene Expression Using Chromatin Features in Various Cellular Contexts." *Genome Biology* 13 (9): R53.
Duren, Zhana, Xi Chen, Rui Jiang, Yong Wang, and Wing Hung Wong. 2017. "Modeling Gene Regulation from Paired Expression and Chromatin Accessibility Data." *Proceedings of the National Academy of Sciences of the United States of America* 114 (25): E4914–23.
Ebina-Shibuya, Risa, Mitsuyo Matsumoto, Makoto Kuwahara, Kyoung-Jin Jang, Manabu Sugai, Yoshiaki Ito, Ryo Funayama, et al. 2017. "Inflammatory Responses Induce an Identity Crisis of Alveolar Macrophages, Leading to Pulmonary Alveolar Proteinosis." *The Journal of Biological Chemistry* 292 (44): 18098–112.
ENCODE Project Consortium. 2012. "An Integrated Encyclopedia of DNA Elements in the Human Genome." *Nature* 489 (7414): 57–74.
Evren, Elza, Emma Ringqvist, and Tim Willinger. 2020. "Origin and Ontogeny of Lung Macrophages: From Mice to Humans." *Immunology* 160 (2): 126–38.
Fan, Xiaoying, Cheng Yang, Wen Li, Xiuzhen Bai, Xin Zhou, Haoling Xie, Lu Wen, and Fuchou Tang. 2021. "SMOOTH-Seq: Single-Cell Genome Sequencing of Human Cells on a Third-Generation Sequencing Platform." *Genome Biology* 22 (1): 195.
Fang, Rongxin, Sebastian Preissl, Yang Li, Xiaomeng Hou, Jacinta Lucero, Xinxin Wang, Amir Motamedi, et al. 2021. "Comprehensive Analysis of Single Cell ATAC-Seq Data with SnapATAC." *Nature Communications* 12 (1): 1–15.
Farbehi, Nona, Ralph Patrick, Aude Dorison, Munira Xaymardan, Vaibhao Janbandhu, Katharina Wystub-Lis, Joshua Wk Ho, Robert E. Nordon, and Richard P. Harvey. 2019. "Single-Cell Expression Profiling Reveals Dynamic Flux of Cardiac Stromal, Vascular and Immune Cells in Health and Injury." *ELife* 8 (March). https://doi.org/10.7554/eLife.43882.
Fardi, Masoumeh, Saeed Solali, and Majid Farshdousti Hagh. 2018. "Epigenetic Mechanisms as a New Approach in Cancer Treatment: An Updated Review." *Genes & Diseases* 5 (4): 304–11.
Fedele, Laura, and Thomas Brand. 2020. "The Intrinsic Cardiac Nervous System and Its Role in Cardiac Pacemaking and Conduction." *Journal of Cardiovascular Development and Disease* 7 (4). https://doi.org/10.3390/jcdd7040054.
Ferrucci, Luigi, and Elisa Fabbri. 2018. "Inflammageing: Chronic Inflammation in Ageing, Cardiovascular Disease, and Frailty." *Nature Reviews. Cardiology* 15 (9): 505–22.
Flick, Kirsten, and Qijun Chen. 2004. "Var Genes, PfEMP1 and the Human Host." *Molecular and Biochemical Parasitology* 134 (1): 3–9.

Flitsch, Lea J., Karen E. Laupman, and Oliver Brüstle. 2020. “Transcription Factor-Based Fate Specification and Forward Programming for Neural Regeneration.” *Frontiers in Cellular Neuroscience* 14 (May): 121.

Gao, Lei, Li-You Wang, Zhi-Qiang Liu, Dan Jiang, Shi-Yong Wu, Yu-Qian Guo, Hong-Mei Tao, et al. 2020. “TNAP Inhibition Attenuates Cardiac Fibrosis Induced by Myocardial Infarction through Deactivating TGF-B1/Smads and Activating P53 Signaling Pathways.” *Cell Death & Disease* 11 (1): 44.

Gates, Leah A., Charles E. Foulds, and Bert W. O’Malley. 2017. “Histone Marks in the ‘Driver’s Seat’: Functional Roles in Steering the Transcription Cycle.” *Trends in Biochemical Sciences* 42 (12): 977–89.

Gawad, Charles, Winston Koh, and Stephen R. Quake. 2016. “Single-Cell Genome Sequencing: Current State of the Science.” *Nature Reviews. Genetics* 17 (3): 175–88.

Gissot, Mathieu, Sylvie Briquet, Philippe Refour, Charlotte Boschet, and Catherine Vaquero. 2005. “PfMyb1, a Plasmodium Falciparum Transcription Factor, Is Required for Intra-Erythrocytic Growth and Controls Key Genes for Cell Cycle Regulation.” *Journal of Molecular Biology* 346 (1): 29–42.

Gogiraju, Rajinikanth, Xingbo Xu, Magdalena L. Bochenek, Julia H. Steinbrecher, Stephan E. Lehnart, Philip Wenzel, Michael Kessel, Elisabeth M. Zeisberg, Matthias Dobbelstein, and Katrin Schäfer. 2015. “Endothelial P53 Deletion Improves Angiogenesis and Prevents Cardiac Fibrosis and Heart Failure Induced by Pressure Overload in Mice.” *Journal of the American Heart Association* 4 (2). https://doi.org/10.1161/JAHA.115.001770.

Gold, David A., Sung Hee Baek, Nicholas J. Schork, David W. Rose, Delaine D. Larsen, Benjamin D. Sachs, Michael G. Rosenfeld, and Bruce A. Hamilton. 2003. “RORalpha Coordinates Reciprocal Signaling in Cerebellar Development through Sonic Hedgehog and Calcium-Dependent Pathways.” *Neuron* 40 (6): 1119–31.

González, Alvaro J., Manu Setty, and Christina S. Leslie. 2015. “Early Enhancer Establishment and Regulatory Locus Complexity Shape Transcriptional Programs in Hematopoietic Differentiation.” *Nature Genetics* 47 (11): 1249–59.

Granja, Jeffrey M., M. Ryan Corces, Sarah E. Pierce, S. Tansu Bagdatli, Hani Choudhry, Howard Y. Chang, and William J. Greenleaf. 2021. “ArchR Is a Scalable Software Package for Integrative Single-Cell Chromatin Accessibility Analysis.” *Nature Genetics* 53 (3): 403–11.

Grant, Charles E., Timothy L. Bailey, and William Stafford Noble. 2011. “FIMO: Scanning for Occurrences of a given Motif.” *Bioinformatics* 27 (7): 1017–18.

Grün, Dominic, Anna Lyubimova, Lennart Kester, Kay Wiebrands, Onur Basak, Nobuo Sasaki, Hans Clevers, and Alexander van Oudenaarden. 2015. “Single-Cell Messenger RNA Sequencing Reveals Rare Intestinal Cell Types.” *Nature* 525 (7568): 251–55.

Guerrini, Valentina, and Maria Laura Gennaro. 2019. “Foam Cells: One Size Doesn’t Fit All.” *Trends in Immunology* 40 (12): 1163–79.

Gupta, Archna P., Wai Hoe Chin, Lei Zhu, Sachel Mok, Yen-Hoon Luah, Eng-How Lim, and Zbynek Bozdech. 2013. “Dynamic Epigenetic Regulation of Gene Expression during the Life Cycle of Malaria Parasite Plasmodium Falciparum.” *PLoS Pathogens* 9 (2): e1003170.

Gusev, Alexander, S. Hong Lee, Gosia Trynka, Hilary Finucane, Bjarni J. Vilhjálmsson, Han Xu, Chongzhi Zang, et al. 2014. “Partitioning Heritability of Regulatory and Cell-Type-

Specific Variants across 11 Common Diseases." *American Journal of Human Genetics* 95 (5): 535–52.
Hadchouel, Alice, David Drummond, Rola Abou Taam, Muriel Lebourgeois, Christophe Delacourt, and Jacques de Blic. 2020. "Alveolar Proteinosis of Genetic Origins." *European Respiratory Review: An Official Journal of the European Respiratory Society*. European Respiratory Society (ERS).
Haghverdi, Laleh, Aaron T. L. Lun, Michael D. Morgan, and John C. Marioni. 2018. "Batch Effects in Single-Cell RNA-Sequencing Data Are Corrected by Matching Mutual Nearest Neighbors." *Nature Biotechnology* 36 (5): 421–27.
Hall, Caitlin, Katja Gehmlich, Chris Denning, and Davor Pavlovic. 2021. "Complex Relationship Between Cardiac Fibroblasts and Cardiomyocytes in Health and Disease." *Journal of the American Heart Association* 10 (5): e019338.
Hamilton, Raymond F., Jr, Sheetal A. Thakur, and Andrij Holian. 2008. "Silica Binding and Toxicity in Alveolar Macrophages." *Free Radical Biology & Medicine* 44 (7): 1246–58.
Han, Xiaoping, Renying Wang, Yincong Zhou, Lijiang Fei, Huiyu Sun, Shujing Lai, Assieh Saadatpour, et al. 2018. "Mapping the Mouse Cell Atlas by Microwell-Seq." *Cell* 172 (5): 1091-1107.e17.
Hart, Rosie, and David R. Greaves. 2010. "Chemerin Contributes to Inflammation by Promoting Macrophage Adhesion to VCAM-1 and Fibronectin through Clustering of VLA-4 and VLA-5." *Journal of Immunology* 185 (6): 3728–39.
Heger, Jacqueline, Julia Bornbaum, Alona Würfel, Christian Hill, Nils Brockmann, Renáta Gáspár, János Pálóczi, et al. 2018. "JDP2 Overexpression Provokes Cardiac Dysfunction in Mice." *Scientific Reports* 8 (1): 7647.
Herzenberg, Leonard A., David Parks, Bita Sahaf, Omar Perez, Mario Roederer, and Leonore A. Herzenberg. 2002. "The History and Future of the Fluorescence Activated Cell Sorter and Flow Cytometry: A View from Stanford." *Clinical Chemistry* 48 (10): 1819–27.
Ho, Karen J., Matthew Spite, Christopher D. Owens, Hope Lancero, Alex H. K. Kroemer, Reena Pande, Mark A. Creager, Charles N. Serhan, and Michael S. Conte. 2010. "Aspirin-Triggered Lipoxin and Resolvin E1 Modulate Vascular Smooth Muscle Phenotype and Correlate with Peripheral Atherosclerosis." *The American Journal of Pathology* 177 (4): 2116–23.
Hocker, James D., Olivier B. Poirion, Fugui Zhu, Justin Buchanan, Kai Zhang, Joshua Chiou, Tsui-Min Wang, et al. 2021. "Cardiac Cell Type-Specific Gene Regulatory Programs and Disease Risk Association." *Science Advances* 7 (20). https://doi.org/10.1126/sciadv.abf1444.
Hosokawa, Hiroyuki, and Ellen V. Rothenberg. 2021. "How Transcription Factors Drive Choice of the T Cell Fate." *Nature Reviews. Immunology* 21 (3): 162–76.
Howe, Kathryn L., and Jason E. Fish. 2019. "Transforming Endothelial Cells in Atherosclerosis." *Nature Metabolism* 1 (9): 856–57.
Huaux, François. 2007. "New Developments in the Understanding of Immunology in Silicosis." *Current Opinion in Allergy and Clinical Immunology* 7 (2): 168–73.
Hulsmans, Maarten, Sebastian Clauss, Ling Xiao, Aaron D. Aguirre, Kevin R. King, Alan Hanley, William J. Hucker, et al. 2017. "Macrophages Facilitate Electrical Conduction in the Heart." *Cell* 169 (3): 510-522.e20.
Ihara, Dai, Yusuke Watanabe, Daiki Seya, Yuji Arai, Yoshie Isomoto, Atsushi Nakano, Atsushi Kubo, Toshihiko Ogura, Teruhisa Kawamura, and Osamu Nakagawa. 2020. "Expression

of Hey2 Transcription Factor in the Early Embryonic Ventricles Is Controlled through a Distal Enhancer by Tbx20 and Gata Transcription Factors." *Developmental Biology* 461 (2): 124–31.

Jablonski, Kyle A., Stephanie A. Amici, Lindsay M. Webb, Juan de Dios Ruiz-Rosado, Phillip G. Popovich, Santiago Partida-Sanchez, and Mireia Guerau-de-Arellano. 2015. "Novel Markers to Delineate Murine M1 and M2 Macrophages." *PloS One* 10 (12): e0145342.

Jaitin, Diego Adhemar, Ephraim Kenigsberg, Hadas Keren-Shaul, Naama Elefant, Franziska Paul, Irina Zaretsky, Alexander Mildner, et al. 2014. "Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types." *Science (New York, N.Y.)* 343 (6172): 776–79.

Jia, Sujie, Shuang Yang, Pei Du, Keqin Gao, Yu Cao, Baige Yao, Ren Guo, and Ming Zhao. 2019. "Regulatory Factor X1 Downregulation Contributes to Monocyte Chemoattractant Protein-1 Overexpression in CD14+ Monocytes via Epigenetic Mechanisms in Coronary Heart Disease." *Frontiers in Genetics* 10 (November): 1098.

Jiang, Lubin, Jianbing Mu, Qingfeng Zhang, Ting Ni, Prakash Srinivasan, Kempaiah Rayavara, Wenjing Yang, et al. 2013. "PfSETvs Methylation of Histone H3K36 Represses Virulence Genes in Plasmodium Falciparum." *Nature* 499 (7457): 223–27.

Jin, Yarong, Ruilei Li, Zhiwei Zhang, Jinjin Ren, Xin Song, and Gong Zhang. 2020. "ZBED1/DREF: A Transcription Factor That Regulates Cell Proliferation." *Oncology Letters* 20 (5): 137.

Jungen, Christiane, Katharina Scherschel, Christian Eickholt, Pawel Kuklik, Niklas Klatt, Nadja Bork, Tim Salzbrunn, et al. 2017. "Disruption of Cardiac Cholinergic Neurons Enhances Susceptibility to Ventricular Arrhythmias." *Nature Communications* 8 (January): 14155.

Kafsack, Björn F. C., Núria Rovira-Graells, Taane G. Clark, Cristina Bancells, Valerie M. Crowley, Susana G. Campino, April E. Williams, et al. 2014. "A Transcriptional Switch Underlies Commitment to Sexual Development in Malaria Parasites." *Nature* 507 (7491): 248–52.

Kajstura, Jan, Narasimman Gurusamy, Barbara Ogórek, Polina Goichberg, Carlos Clavo-Rondon, Toru Hosoda, Domenico D'Amario, et al. 2010. "Myocyte Turnover in the Aging Human Heart." *Circulation Research* 107 (11): 1374–86.

Karlić, Rosa, Ho-Ryun Chung, Julia Lasserre, Kristian Vlahovicek, and Martin Vingron. 2010. "Histone Modification Levels Are Predictive for Gene Expression." *Proceedings of the National Academy of Sciences of the United States of America* 107 (7): 2926–31.

Kelley, David R., Yakir A. Reshef, Maxwell Bileschi, David Belanger, Cory Y. McLean, and Jasper Snoek. 2018. "Sequential Regulatory Activity Prediction across Chromosomes with Convolutional Neural Networks." *Genome Research* 28 (5): 739–50.

Kessler, Elise L., Mathilde R. Rivaud, Marc A. Vos, and Toon A. B. van Veen. 2019. "Sex-Specific Influence on Cardiac Structural Remodeling and Therapy in Cardiovascular Disease." *Biology of Sex Differences* 10 (1): 7.

Koenig, Andrew L., Irina Shchukina, Junedh Amrute, Prabhakar S. Andhey, Konstantin Zaitsev, Lulu Lai, Geetika Bajpai, et al. 2022. "Single-Cell Transcriptomics Reveals Cell-Type-Specific Diversification in Human Heart Failure." *Nature Cardiovascular Research* 1 (3): 263–80.

Korhonen, Janne, Petri Martinmäki, Cinzia Pizzi, Pasi Rastas, and Esko Ukkonen. 2009. "MOODS: Fast Search for Position Weight Matrix Matches in DNA Sequences." *Bioinformatics* 25 (23): 3181–82.

Korsunsky, Ilya, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-Ru Loh, and Soumya Raychaudhuri. 2019. “Fast, Sensitive and Accurate Integration of Single-Cell Data with Harmony.” *Nature Methods* 16 (12): 1289–96.

Kosmas, Constantine E., Delia Silverio, Andreas Sourlas, Peter D. Montan, Eliscer Guzman, and Mario J. Garcia. 2019. “Anti-Inflammatory Therapy for Cardiovascular Disease.” *Annals of Translational Medicine* 7 (7): 147.

Kostopoulos, Christos G., Sofia G. Spiroglou, John N. Varakis, Efstratios Apostolakis, and Helen H. Papadaki. 2014. “Chemerin and CMKLR1 Expression in Human Arteries and Periadventitial Fat: A Possible Role for Local Chemerin in Atherosclerosis?” *BMC Cardiovascular Disorders* 14 (April): 56.

Kundaje, Anshul, Manuel Middendorf, Mihir Shah, Chris H. Wiggins, Yoav Freund, and Christina Leslie. 2006. “A Classification-Based Framework for Predicting and Analyzing Gene Regulatory Response.” *BMC Bioinformatics* 7 Suppl 1 (S1): S5.

Lamb, Justin, Emily D. Crawford, David Peck, Joshua W. Modell, Irene C. Blat, Matthew J. Wrobel, Jim Lerner, et al. 2006. “The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease.” *Science (New York, N.Y.)* 313 (5795): 1929–35.

Lee, Bum-Kyu, Akshay A. Bhinge, Anna Battenhouse, Ryan M. McDaniell, Zheng Liu, Lingyun Song, Yunyun Ni, et al. 2012. “Cell-Type Specific and Combinatorial Usage of Diverse Transcription Factors Revealed by Genome-Wide Binding Studies in Multiple Human Cells.” *Genome Research* 22 (1): 9–24.

Leung, Chi Chiu, Ignatius Tak Sun Yu, and Weihong Chen. 2012. “Silicosis.” *Lancet* 379 (9830): 2008–18.

Li, Gang. 2021. “Abstract 12513: Post-GWAS Functional Analysis of the CDKN2A/B Locus Identifies CUX1 as a Regulator of Endothelial Senescence by Modulating the CAD-Associated P16INK4A Expression.” *Circulation* 144 (Suppl_1): A12513–A12513.

Li, Heng, and Richard Durbin. 2010. “Fast and Accurate Long-Read Alignment with Burrows-Wheeler Transform.” *Bioinformatics (Oxford, England)* 26 (5): 589–95.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics (Oxford, England)* 25 (16): 2078–79.

Li, Tian, Xinyu Yang, Hong Xu, and Heliang Liu. 2022. “Early Identification, Accurate Diagnosis, and Treatment of Silicosis.” *Canadian Respiratory Journal: Journal of the Canadian Thoracic Society* 2022 (April): 3769134.

Liberzon, Arthur, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. 2015. “The Molecular Signatures Database (MSigDB) Hallmark Gene Set Collection.” *Cell Systems* 1 (6): 417–25.

Lin, Shin, and Yiing Lin. 1970. “Protocol for Flash Freezing Tissue Sample.” Protocols.Io. January 1, 1970. https://www.protocols.io/view/protocol-for-flash-freezing-tissue-sample-kxygxednkv8j/v1.

———. 2020. “Protocol for Tissue Collection from Organ Procurement Organization V1.” *Protocols.Io*. ZappyLab, Inc. https://doi.org/10.17504/protocols.io.biqpkdvn.

“Linear Mixed-Effects Models Using ‘Eigen’ and S4 [R Package Lme4 Version 1.1-28].” 2022, February. https://cran.r-project.org/web/packages/lme4/index.html.

Litviňuková, Monika, Carlos Talavera-López, Henrike Maatz, Daniel Reichart, Catherine L. Worth, Eric L. Lindberg, Masatoshi Kanda, et al. 2020. “Cells of the Adult Human Heart.” *Nature* 588 (7838): 466–72.

Lopatin, Yury. 2015. “Metabolic Therapy in Heart Failure.” *Cardiac Failure Review* 1 (2): 112–17.

Lopez-Rubio, Jose-Juan, Liliana Mancio-Silva, and Artur Scherf. 2009. “Genome-Wide Analysis of Heterochromatin Associates Clonally Variant Gene Regulation with Perinuclear Repressive Centers in Malaria Parasites.” *Cell Host & Microbe* 5 (2): 179–90.

Lu, Xueqing Maggie, Gayani Batugedara, Michael Lee, Jacques Prudhomme, Evelien M. Bunnik, and Karine G. Le Roch. 2017. “Nascent RNA Sequencing Reveals Mechanisms of Gene Regulation in the Human Malaria Parasite Plasmodium Falciparum.” *Nucleic Acids Research* 45 (13): 7825–40.

Lu, Yang Young, Yingying Fan, Jinchi Lv, and William Stafford Noble. 2018. “DeepPINK: Reproducible Feature Selection in Deep Neural Networks.” *ArXiv [Cs.LG]*. arXiv. https://proceedings.neurips.cc/paper/2018/file/29daf9442f3c0b60642b14c081b4a556-Paper.pdf.

Luisetti, Maurizio, Pierdonato Bruno, Zamir Kadija, Takuji Suzuki, Salvatore Raffa, Maria Rosaria Torrisi, Ilaria Campo, et al. 2011. “Relationship between Diffuse Pulmonary Fibrosis, Alveolar Proteinosis, and Granulocyte-Macrophage Colony Stimulating Factor Autoantibodies.” *Respiratory Care* 56 (10): 1608–10.

Lundberg, Scott, and Su-In Lee. 2017. “A Unified Approach to Interpreting Model Predictions.” *ArXiv [Cs.AI]*. arXiv. http://arxiv.org/abs/1705.07874.

MacQuarrie, Kyle L., Abraham P. Fong, Randall H. Morse, and Stephen J. Tapscott. 2011. “Genome-Wide Transcription Factor Binding: Beyond Direct Target Regulation.” *Trends in Genetics: TIG* 27 (4): 141–48.

Mak, Tak W., Ludger Hauck, Daniela Grothe, and Filio Billia. 2017. “P53 Regulates the Cardiac Transcriptome.” *Proceedings of the National Academy of Sciences of the United States of America* 114 (9): 2331–36.

Martin, Beth K., Chengxiang Qiu, Eva Nichols, Melissa Phung, Rula Green-Gladden, Sanjay Srivatsan, Ronnie Blecher-Gonen, et al. 2021. “An Optimized Protocol for Single Cell Transcriptional Profiling by Combinatorial Indexing.” *ArXiv [q-Bio.GN]*. arXiv. http://arxiv.org/abs/2110.15400.

Martin, Bryan D., Daniela Witten, and Amy D. Willis. 2020. “Modeling Microbial Abundances and Dysbiosis with Beta-Binomial Regression.” *The Annals of Applied Statistics* 14 (1): 94–115.

Mavrich, Travis N., Cizhong Jiang, Ilya P. Ioshikhes, Xiaoyong Li, Bryan J. Venters, Sara J. Zanton, Lynn P. Tomsho, et al. 2008. “Nucleosome Organization in the Drosophila Genome.” *Nature* 453 (7193): 358–62.

McInnes, Leland, John Healy, and James Melville. 2018. “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.” *ArXiv [Stat.ML]*. arXiv. http://arxiv.org/abs/1802.03426.

McLeay, Robert C., Tom Lesluyes, Gabriel Cuellar Partida, and Timothy L. Bailey. 2012. “Genome-Wide in Silico Prediction of Gene Expression.” *Bioinformatics (Oxford, England)* 28 (21): 2789–96.

Mehta, Amit K., Donald T. Gracias, and Michael Croft. 2018. “TNF Activity and T Cells.” *Cytokine* 101 (January): 14–18.

Meng, Xiao-Ming, David J. Nikolic-Paterson, and Hui Yao Lan. 2016. “TGF-β: The Master Regulator of Fibrosis.” *Nature Reviews. Nephrology* 12 (6): 325–38.
Middendorf, Manuel, Anshul Kundaje, Chris Wiggins, Yoav Freund, and Christina Leslie. 2004. “Predicting Genetic Regulatory Response Using Classification.” *Bioinformatics (Oxford, England)* 20 Suppl 1 (Suppl 1): i232-40.
Muers, Mary. 2011. “Functional Genomics: The ModENCODE Guide to the Genome.” *Nature Reviews. Genetics*. Springer Science and Business Media LLC.
Mullen, Alan C., David A. Orlando, Jamie J. Newman, Jakob Lovén, Roshan M. Kumar, Steve Bilodeau, Jessica Reddy, Matthew G. Guenther, Rodney P. DeKoter, and Richard A. Young. 2011. “Master Transcription Factors Determine Cell-Type-Specific Responses to TGF-β Signaling.” *Cell* 147 (3): 565–76.
Murphy, Sean P., Rahul Kakkar, Cian P. McCarthy, and James L. Januzzi Jr. 2020. “Inflammation in Heart Failure: JACC State-of-the-Art Review.” *Journal of the American College of Cardiology* 75 (11): 1324–40.
Nakerakanti, Sashidhar S., Andreea M. Bujor, and Maria Trojanowska. 2011. “CCN2 Is Required for the TGF-β Induced Activation of Smad1-Erk1/2 Signaling Network.” *PloS One* 6 (7): e21911.
Narlikar, Leelavati, Noboru J. Sakabe, Alexander A. Blanski, Fabio E. Arimura, John M. Westlund, Marcelo A. Nobrega, and Ivan Ovcharenko. 2010. “Genome-Wide Discovery of Human Heart Enhancers.” *Genome Research* 20 (3): 381–92.
Oeckinghaus, Andrea, and Sankar Ghosh. 2009. “The NF-KappaB Family of Transcription Factors and Its Regulation.” *Cold Spring Harbor Perspectives in Biology* 1 (4): a000034.
Oliva, Meritxell, Manuel Muñoz-Aguirre, Sarah Kim-Hellmuth, Valentin Wucher, Ariel D. H. Gewirtz, Daniel J. Cotter, Princy Parsana, et al. 2020. “The Impact of Sex on Gene Expression across Human Tissues.” *Science* 369 (6509). https://doi.org/10.1126/science.aba3066.
Olivetti, G., G. Giordano, D. Corradi, M. Melissari, C. Lagrasta, S. R. Gambert, and P. Anversa. 1995. “Gender Differences and Aging: Effects on the Human Heart.” *Journal of the American College of Cardiology* 26 (4): 1068–79.
Olivetti, G., M. Melissari, J. M. Capasso, and P. Anversa. 1991. “Cardiomyopathy of the Aging Human Heart. Myocyte Loss and Reactive Cellular Hypertrophy.” *Circulation Research* 68 (6): 1560–68.
Osmanbeyoglu, Hatice U., Fumiko Shimizu, Angela Rynne-Vidal, Direna Alonso-Curbelo, Hsuan-An Chen, Hannah Y. Wen, Tsz-Lun Yeung, et al. 2019. “Chromatin-Informed Inference of Transcriptional Programs in Gynecologic and Basal Breast Cancers.” *Nature Communications* 10 (1): 4369.
Ouyang, Zhengqing, Qing Zhou, and Wing Hung Wong. 2009. “ChIP-Seq of Transcription Factors Predicts Absolute and Differential Gene Expression in Embryonic Stem Cells.” *Proceedings of the National Academy of Sciences of the United States of America* 106 (51): 21521–26.
Painter, Heather J., Tracey L. Campbell, and Manuel Llinás. 2011. “The Apicomplexan AP2 Family: Integral Factors Regulating Plasmodium Development.” *Molecular and Biochemical Parasitology* 176 (1): 1–7.
Parichatikanond, Warisara, Theerut Luangmonkong, Supachoke Mangmool, and Hitoshi Kurose. 2020. “Therapeutic Targets for the Treatment of Cardiac Fibrosis and Cancer: Focusing on TGF-β Signaling.” *Frontiers in Cardiovascular Medicine* 7 (March): 34.

Park, Jin Hee, Na Kyung Lee, and Soo Young Lee. 2017. “Current Understanding of RANK Signaling in Osteoclast Differentiation and Maturation.” *Molecules and Cells* 40 (10): 706–13.

Pe’er, Dana, Aviv Regev, and Amos Tanay. 2002. “Minreg: Inferring an Active Regulator Set.” *Bioinformatics (Oxford, England)* 18 Suppl 1: S258-67.

Petter, Michaela, Shamista A. Selvarajah, Chin Chin Lee, Wai Hoe Chin, Archna P. Gupta, Zbynek Bozdech, Graham V. Brown, and Michael F. Duffy. 2013. “H2A.Z and H2B.Z Double-Variant Nucleosomes Define Intergenic Regions and Dynamically Occupy Var Gene Promoters in the Malaria Parasite Plasmodium Falciparum.” *Molecular Microbiology* 87 (6): 1167–82.

Picelli, Simone, Omid R. Faridani, Asa K. Björklund, Gösta Winberg, Sven Sagasser, and Rickard Sandberg. 2014. “Full-Length RNA-Seq from Single Cells Using Smart-Seq2.” *Nature Protocols* 9 (1): 171–81.

Pliner, Hannah A., A. Gogate, and B. Ewing. n.d. “Bb-Lab/Bb-Sci.” Github. Accessed February 1, 2020. https://github.com/bbi-lab/bbi-sci.

Pliner, Hannah A., Jonathan S. Packer, José L. McFaline-Figueroa, Darren A. Cusanovich, Riza M. Daza, Delasa Aghamirzaie, Sanjay Srivatsan, et al. 2018. “Cicero Predicts Cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data.” *Molecular Cell* 71 (5): 858-871.e8.

Ponts, Nadia, Elena Y. Harris, Jacques Prudhomme, Ivan Wick, Colleen Eckhardt-Ludka, Glenn R. Hicks, Gary Hardiman, Stefano Lonardi, and Karine G. Le Roch. 2010. “Nucleosome Landscape and Control of Transcription in the Human Malaria Parasite.” *Genome Research* 20 (2): 228–38.

Prat, Yosef, Menachem Fromer, Nathan Linial, and Michal Linial. 2011. “Recovering Key Biological Constituents through Sparse Representation of Gene Expression.” *Bioinformatics (Oxford, England)* 27 (5): 655–61.

Programme, Global Malaria. 2017. “World Malaria Report 2017.” World Health Organization. November 19, 2017. https://www.who.int/publications/i/item/9789241565523.

Puvogel, Sofía, Astrid Alsema, Laura Kracht, Maree J. Webster, Cynthia Shannon Weickert, Iris E. C. Sommer, and Bart J. L. Eggen. 2022. “Single-Nucleus RNA Sequencing of Midbrain Blood-Brain Barrier Cells in Schizophrenia Reveals Subtle Transcriptional Changes with Overall Preservation of Cellular Proportions and Phenotypes.” *Molecular Psychiatry*, October, 1–10.

Quinlan, Aaron R., and Ira M. Hall. 2010. “BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features.” *Bioinformatics (Oxford, England)* 26 (6): 841–42.

Read, David F., Gregory T. Booth, Riza M. Daza, Dana L. Jackson, Rula Green Gladden, Sanjay R. Srivatsan, Brent Ewing, et al. 2022. “Single-Cell Analysis of Chromatin and Expression Reveals Age- and Sex-Associated Alterations in the Human Heart.” *BioRxiv*. https://doi.org/10.1101/2022.07.12.496461.

Read, David F., Kate Cook, Yang Y. Lu, Karine G. Le Roch, and William Stafford Noble. 2019. “Predicting Gene Expression in the Human Malaria Parasite Plasmodium Falciparum Using Histone Modification, Nucleosome Positioning, and 3D Localization Features.” *PLoS Computational Biology* 15 (9): e1007329.

Redfield, Margaret M., Steven J. Jacobsen, Barry A. Borlaug, Richard J. Rodeheffer, and David A. Kass. 2005. “Age- and Gender-Related Ventricular-Vascular Stiffening: A Community-Based Study.” *Circulation* 112 (15): 2254–62.

Reichart, Daniel, Eric L. Lindberg, Henrike Maatz, Antonio M. A. Miranda, Anissa Viveiros, Nikolay Shvetsov, Anna Gärtner, et al. 2022. “Pathogenic Variants Damage Cell Composition and Single Cell Transcription in Cardiomyopathies.” *Science* 377 (6606): eabo1984.

Reyes-Palomares, Armando, Mingxia Gu, Fabian Grubert, Ivan Berest, Silin Sa, Maya Kasowski, Christian Arnold, et al. 2020. “Remodeling of Active Endothelial Enhancers Is Associated with Aberrant Gene-Regulatory Networks in Pulmonary Arterial Hypertension.” *Nature Communications* 11 (1): 1673.

Rhie, Arang, Shane A. McCarthy, Olivier Fedrigo, Joana Damas, Giulio Formenti, Sergey Koren, Marcela Uliano-Silva, et al. 2021. “Towards Complete and Error-Free Genome Assemblies of All Vertebrate Species.” *Nature* 592 (7856): 737–46.

Riechmann, J. L., and E. M. Meyerowitz. 1998. “The AP2/EREBP Family of Plant Transcription Factors.” *Biological Chemistry* 379 (6): 633–46.

Ritchie, Matthew E., Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. 2015. “Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies.” *Nucleic Acids Research* 43 (7): e47.

Robb, L., C. C. Drinkwater, D. Metcalf, R. Li, F. Köntgen, N. A. Nicola, and C. G. Begley. 1995. “Hematopoietic and Lung Abnormalities in Mice with a Null Mutation of the Common Beta Subunit of the Receptors for Granulocyte-Macrophage Colony-Stimulating Factor and Interleukins 3 and 5.” *Proceedings of the National Academy of Sciences of the United States of America* 92 (21): 9565–69.

Robin, Xavier, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. 2011. “PROC: An Open-Source Package for R and S+ to Analyze and Compare ROC Curves.” *BMC Bioinformatics* 12 (1): 77.

Rodgers, Jennifer L., Jarrod Jones, Samuel I. Bolleddu, Sahit Vanthenapalli, Lydia E. Rodgers, Kinjal Shah, Krishna Karia, and Siva K. Panguluri. 2019. “Cardiovascular Risks Associated with Gender and Aging.” *Journal of Cardiovascular Development and Disease* 6 (2): 19.

Rodríguez-Enríquez, Sara, Álvaro Marín-Hernández, Juan Carlos Gallardo-Pérez, Silvia Cecilia Pacheco-Velázquez, Javier Alejandro Belmont-Díaz, Diana Xochiquetzal Robledo-Cadena, Jorge Luis Vargas-Navarro, Norma Angélica Corona de la Peña, Emma Saavedra, and Rafael Moreno-Sánchez. 2019. “Transcriptional Regulation of Energy Metabolism in Cancer Cells.” *Cells* 8 (10). https://doi.org/10.3390/cells8101225.

Ross, Sarah H., and Doreen A. Cantrell. 2018. “Signaling and Function of Interleukin-2 in T Lymphocytes.” *Annual Review of Immunology* 36 (April): 411–33.

Rufaihah, Abdul Jalil, Husnain Khawaja Haider, Boon Chin Heng, Lei Ye, Ru San Tan, Wei Seong Toh, Xian Feng Tian, Eugene Kwang-Wei Sim, and Tong Cao. 2010. “Therapeutic Angiogenesis by Transplantation of Human Embryonic Stem Cell-Derived CD133+ Endothelial Progenitor Cells for Cardiac Repair.” *Regenerative Medicine* 5 (2): 231–44.

Ruiz-Villalba, Adrián, Juan P. Romero, Silvia C. Hernández, Amaia Vilas-Zornoza, Nikolaus Fortelny, Laura Castro-Labrador, Patxi San Martin-Uriz, et al. 2020. “Single-Cell RNA Sequencing Analysis Reveals a Crucial Role for CTHRC1 (Collagen Triple Helix Repeat Containing 1) Cardiac Fibroblasts After Myocardial Infarction.” *Circulation* 142 (19): 1831–47.

Ruzicka, W. Brad, Shahin Mohammadi, Jose Davila-Velderrain, Sivan Subburaju, Daniel Reed Tso, Makayla Hourihan, and Manolis Kellis. 2020. “Single-Cell Dissection of Schizophrenia Reveals Neurodevelopmental-Synaptic Axis and Transcriptional Resilience.” *BioRxiv*. medRxiv. https://doi.org/10.1101/2020.11.06.20225342.

Sandelin, Albin, Wynand Alkema, Pär Engström, Wyeth W. Wasserman, and Boris Lenhard. 2004. “JASPAR: An Open-Access Database for Eukaryotic Transcription Factor Binding Profiles.” *Nucleic Acids Research* 32 (Database issue): D91-4.

Santos, Joana Mendonca, Gabrielle Josling, Philipp Ross, Preeti Joshi, Lindsey Orchard, Tracey Campbell, Ariel Schieler, Ileana M. Cristea, and Manuel Llinás. 2017. “Red Blood Cell Invasion by the Malaria Parasite Is Coordinated by the PfAP2-I Transcription Factor.” *Cell Host & Microbe* 21 (6): 731-741.e10.

Satpathy, Ansuman T., Jeffrey M. Granja, Kathryn E. Yost, Yanyan Qi, Francesca Meschi, Geoffrey P. McDermott, Brett N. Olsen, et al. 2019. “Massively Parallel Single-Cell Chromatin Landscapes of Human Immune Cell Development and Intratumoral T Cell Exhaustion.” *Nature Biotechnology* 37 (8): 925–36.

Schoof, Erwin M., Benjamin Furtwängler, Nil Üresin, Nicolas Rapin, Simonas Savickas, Coline Gentil, Eric Lechman, Ulrich Auf Dem Keller, John E. Dick, and Bo T. Porse. 2021. “Quantitative Single-Cell Proteomics as a Tool to Characterize Cellular Hierarchies.” *Nature Communications* 12 (1): 3341.

Segal, E., R. Yelensky, and D. Koller. 2003. “Genome-Wide Discovery of Transcriptional Modules from DNA Sequence and Gene Expression.” *Bioinformatics (Oxford, England)* 19 Suppl 1: i273-82.

Sekine, Ayumi, Tetsu Nishiwaki, Rintaro Nishimura, Takeshi Kawasaki, Takashi Urushibara, Rika Suda, Toshio Suzuki, et al. 2016. “Prominin-1/CD133 Expression as Potential Tissue-Resident Vascular Endothelial Progenitor Cells in the Pulmonary Circulation.” *American Journal of Physiology. Lung Cellular and Molecular Physiology* 310 (11): L1130-42.

Sergushichev, Alexey A. 2016. “An Algorithm for Fast Preranked Gene Set Enrichment Analysis Using Cumulative Statistic Calculation.” *BioRxiv*. https://doi.org/10.1101/060012.

Seydel, Caroline. 2021. “Single-Cell Metabolomics Hits Its Stride.” *Nature Methods* 18 (12): 1452–56.

Shen, Mark J., and Douglas P. Zipes. 2014. “Role of the Autonomic Nervous System in Modulating Cardiac Arrhythmias.” *Circulation Research* 114 (6): 1004–21.

Shibata, Y., P. Y. Berclaz, Z. C. Chroneos, M. Yoshida, J. A. Whitsett, and B. C. Trapnell. 2001. “GM-CSF Regulates Alveolar Macrophage Differentiation and Innate Immunity in the Lung through PU.1.” *Immunity* 15 (4): 557–67.

Shima, Kenjiro, Paritha Arumugam, Yan Ma, Dianna Black, Claudia Chalk, Brenna Carey, and Bruce C. Trapnell. n.d. “Development and Validation of Csf2ra Gene-Deficient Mice as a Clinically Relevant Model of Children with Hereditary Pulmonary Alveolar Proteinosis.” In *B108. CYSTIC FIBROSIS, PRIMARY CILIARY DYSKINESIA, AND ILD*, A4837–A4837.

Shimizu, Ippei, and Tohru Minamino. 2019. “Cellular Senescence in Cardiac Diseases.” *Journal of Cardiology* 74 (4): 313–19.

Shirai, Tsuyoshi, Marc Hilhorst, David G. Harrison, Jörg J. Goronzy, and Cornelia M. Weyand. 2015. “Macrophages in Vascular Inflammation--From Atherosclerosis to Vasculitis.” *Autoimmunity* 48 (3): 139–51.

Singh, Ritambhara, Jack Lanchantin, Gabriel Robins, and Yanjun Qi. 2016. “DeepChrome: Deep-Learning for Predicting Gene Expression from Histone Modifications.” *Bioinformatics* 32 (17): i639–48.

Squair, Jordan W., Matthieu Gautier, Claudia Kathe, Mark A. Anderson, Nicholas D. James, Thomas H. Hutson, Rémi Hudelle, et al. 2021. “Confronting False Discoveries in Single-Cell Differential Expression.” *Nature Communications* 12 (1): 5692.

Stanley, E., G. J. Lieschke, D. Grail, D. Metcalf, G. Hodgson, J. A. Gall, D. W. Maher, J. Cebon, V. Sinickas, and A. R. Dunn. 1994. “Granulocyte/Macrophage Colony-Stimulating Factor-Deficient Mice Show No Major Perturbation of Hematopoiesis but Develop a Characteristic Pulmonary Pathology.” *Proceedings of the National Academy of Sciences of the United States of America* 91 (12): 5592–96.

Stark, Rory, Marta Grzelak, and James Hadfield. 2019. “RNA Sequencing: The Teenage Years.” *Nature Reviews. Genetics* 20 (11): 631–56.

Steenman, Marja, and Gilles Lande. 2017. “Cardiac Aging and Heart Disease in Humans.” *Biophysical Reviews* 9 (2): 131–37.

Stergiopoulos, Athanasios, Maximilianos Elkouris, and Panagiotis K. Politis. 2014. “Prospero-Related Homeobox 1 (Prox1) at the Crossroads of Diverse Pathways during Adult Neural Fate Specification.” *Frontiers in Cellular Neuroscience* 8: 454.

Strunz, Maximilian, Lukas M. Simon, Meshal Ansari, Jaymin J. Kathiriya, Ilias Angelidis, Christoph H. Mayr, George Tsidiridis, et al. 2020. “Alveolar Regeneration through a Krt8+ Transitional Stem Cell State That Persists in Human Lung Fibrosis.” *Nature Communications* 11 (1): 3559.

Subramanian, Aravind, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, et al. 2005. “Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (43): 15545–50.

Sugiaman-Trapman, Debora, Morana Vitezic, Eeva-Mari Jouhilahti, Anthony Mathelier, Gilbert Lauter, Sougat Misra, Carsten O. Daub, Juha Kere, and Peter Swoboda. 2018. “Characterization of the Human RFX Transcription Factor Family by Regulatory and Target Gene Analysis.” *BMC Genomics* 19 (1): 181.

Suzuki, T., B. Maranda, T. Sakagami, P. Catellier, C-Y Couture, B. C. Carey, C. Chalk, and B. C. Trapnell. 2011. “Hereditary Pulmonary Alveolar Proteinosis Caused by Recessive CSF2RB Mutations.” *The European Respiratory Journal: Official Journal of the European Society for Clinical Respiratory Physiology* 37 (1): 201–4.

Suzuki, Takuji, Paritha Arumugam, Takuro Sakagami, Nico Lachmann, Claudia Chalk, Anthony Sallese, Shuichi Abe, et al. 2014. “Pulmonary Macrophage Transplantation Therapy.” *Nature* 514 (7523): 450–54.

Tabula Muris Consortium. 2020. “A Single-Cell Transcriptomic Atlas Characterizes Ageing Tissues in the Mouse.” *Nature* 583 (7817): 590–95.

Tabula Muris Consortium, Overall coordination, Logistical coordination, Organ collection and processing, Library preparation and sequencing, Computational data analysis, Cell type annotation, Writing group, Supplemental text writing group, and Principal investigators.

2018. “Single-Cell Transcriptomics of 20 Mouse Organs Creates a Tabula Muris.” *Nature* 562 (7727): 367–72.

Thakar, Amit, Pooja Gupta, William T. McAllister, and Jordanka Zlatanova. 2010. “Histone Variant H2A.Z Inhibits Transcription in Reconstituted Nucleosomes.” *Biochemistry* 49 (19): 4018–26.

Thakur, Sheetal A., Raymond Hamilton Jr, Timo Pikkarainen, and Andrij Holian. 2009. “Differential Binding of Inorganic Particles to MARCO.” *Toxicological Sciences: An Official Journal of the Society of Toxicology* 107 (1): 238–46.

Toenhake, Christa Geeke, Sabine Anne-Kristin Fraschka, Mahalingam Shanmugiah Vijayabaskar, David Robert Westhead, Simon Jan van Heeringen, and Richárd Bártfai. 2018. “Chromatin Accessibility-Based Characterization of the Gene Regulatory Network Underlying Plasmodium Falciparum Blood-Stage Development.” *Cell Host & Microbe* 23 (4): 557-569.e9.

Traag, V. A., L. Waltman, and N. J. van Eck. 2019. “From Louvain to Leiden: Guaranteeing Well-Connected Communities.” *Scientific Reports* 9 (1): 5233.

Trapnell, Bruce C., Koh Nakata, Francesco Bonella, Ilaria Campo, Matthias Griese, John Hamilton, Tisha Wang, Cliff Morgan, Vincent Cottin, and Cormac McCarthy. 2019. “Pulmonary Alveolar Proteinosis.” *Nature Reviews. Disease Primers* 5 (1): 16.

Trapnell, Cole, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J. Lennon, Kenneth J. Livak, Tarjei S. Mikkelsen, and John L. Rinn. 2014. “The Dynamics and Regulators of Cell Fate Decisions Are Revealed by Pseudotemporal Ordering of Single Cells.” *Nature Biotechnology* 32 (4): 381–86.

Tsai, Ming-Ho, Kenly Wuputra, Yin-Chu Lin, Chang-Shen Lin, and Kazunari K. Yokoyama. 2016. “Multiple Functions of the Histone Chaperone Jun Dimerization Protein 2.” *Gene* 590 (2): 193–200.

Tual-Chalot, Simon, Marwa Mahmoud, Kathleen R. Allinson, Rachael E. Redgrave, Zhenhua Zhai, S. Paul Oh, Marcus Fruttiger, and Helen M. Arthur. 2014. “Endothelial Depletion of Acvrl1 in Mice Leads to Arteriovenous Malformations Associated with Reduced Endoglin Expression.” *PloS One* 9 (6): e98646.

Tucker, Nathan R., Mark Chaffin, Stephen J. Fleming, Amelia W. Hall, Victoria A. Parsons, Kenneth C. Bedi Jr, Amer-Denis Akkad, et al. 2020. “Transcriptional and Cellular Diversity of the Human Heart.” *Circulation* 142 (5): 466–82.

Uehara, Yasuaki, Nikolaos M. Nikolaidis, Lori B. Pitstick, Huixing Wu, Jane J. Yu, Erik Zhang, Yoshihiro Hasegawa, et al. 2021. “Novel Insights into Pulmonary Phosphate Homeostasis and Osteoclastogenesis Emerge from the Study of Pulmonary Alveolar Microlithiasis.” *BioRxiv*. bioRxiv. https://doi.org/10.1101/2021.07.11.451970.

Varoquaux, Nelle, Ferhat Ay, William Stafford Noble, and Jean-Philippe Vert. 2014. “A Statistical Approach for Inferring the 3D Structure of the Genome.” *Bioinformatics (Oxford, England)* 30 (12): i26-33.

Wagner, G. R. 1997. “Asbestosis and Silicosis.” *Lancet* 349 (9061): 1311–15.

Wang, Bin, Shiju Chen, Hongyan Qian, Rongjuan Chen, Yan He, Xinwei Zhang, Jingxiu Xuan, Yuan Liu, and Guixiu Shi. 2020. “Development and Validation of a Transcriptional Signature for the Assessment of Fibrosis in Organs.” *BioRxiv*. medRxiv. https://doi.org/10.1101/2020.03.14.20024141.

Wang, Yan, and Ji-Guang Wang. 2019. “Genome-Wide Association Studies of Hypertension and Several Other Cardiovascular Diseases.” *The Pulse of the Montana State Nurses' Association* 6 (3–4): 169–86.

Weber, Christopher M., Jorja G. Henikoff, and Steven Henikoff. 2010. “H2A.Z Nucleosomes Enriched over Active Genes Are Homotypic.” *Nature Structural & Molecular Biology* 17 (12): 1500–1507.

Weirauch, Matthew T., Ally Yang, Mihai Albu, Atina G. Cote, Alejandro Montenegro-Montero, Philipp Drewe, Hamed S. Najafabadi, et al. 2014. “Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity.” *Cell* 158 (6): 1431–43.

Wolock, Samuel L., Romain Lopez, and Allon M. Klein. 2019. “Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data.” *Cell Systems* 8 (4): 281-291.e9.

Wotton, D., R. S. Lo, S. Lee, and J. Massagué. 1999. “A Smad Transcriptional Corepressor.” *Cell* 97 (1): 29–39.

Xiang, Fan, Yasuhiko Sakata, Lei Cui, Joey M. Youngblood, Hironori Nakagami, James K. Liao, Ronglih Liao, and Michael T. Chin. 2006. “Transcription Factor CHF1/Hey2 Suppresses Cardiac Hypertrophy through an Inhibitory Interaction with GATA4.” *American Journal of Physiology. Heart and Circulatory Physiology* 290 (5): H1997-2006.

Xu, Jian, Samy Lamouille, and Rik Derynck. 2009. “TGF-Beta-Induced Epithelial to Mesenchymal Transition.” *Cell Research* 19 (2): 156–72.

Xu, Mingcong, Xuefeng Bai, Bo Ai, Guorui Zhang, Chao Song, Jun Zhao, Yuezhu Wang, et al. 2022. “TF-Marker: A Comprehensive Manually Curated Database for Transcription Factors and Related Markers in Specific Cell and Tissue Types in Human.” *Nucleic Acids Research* 50 (D1): D402–12.

Young, Matthew D., and Sam Behjati. 2020. “SoupX Removes Ambient RNA Contamination from Droplet-Based Single-Cell RNA Sequencing Data.” *GigaScience* 9 (12). https://doi.org/10.1093/gigascience/giaa151.

Yuda, Masao, Shiroh Iwanaga, Izumi Kaneko, and Tomomi Kato. 2015. “Global Transcriptional Repression: An Initial and Essential Step for Plasmodium Sexual Development.” *Proceedings of the National Academy of Sciences of the United States of America* 112 (41): 12824–29.

Zeng, Wanwen, Yong Wang, and Rui Jiang. 2020. “Integrating Distal and Proximal Information to Predict Gene Expression via a Densely Connected Convolutional Neural Network.” *Bioinformatics* 36 (2): 496–503.

Zhang, Kai, James D. Hocker, Michael Miller, Xiaomeng Hou, Joshua Chiou, Olivier B. Poirion, Yunjiang Qiu, et al. 2021. “A Single-Cell Atlas of Chromatin Accessibility in the Human Genome.” *Cell* 184 (24): 5985-6001.e19.

Zhang, Martin Jinye, Angela Oliveira Pisco, Spyros Darmanis, and James Zou. 2021. “Mouse Aging Cell Atlas Analysis Reveals Global and Cell Type-Specific Aging Signatures.” *ELife* 10 (April): e62293.

Zhang, Xiufeng, Xiaoming Wu, Wenru Tang, and Ying Luo. 2012. “Loss of P16(Ink4a) Function Rescues Cellular Senescence Induced by Telomere Dysfunction.” *International Journal of Molecular Sciences* 13 (5): 5866–77.

Zhang, Yiqiang, Nuria Gago-Lopez, Ning Li, Zhenhe Zhang, Naima Alver, Yonggang Liu, Amy M. Martinson, Avin Mehri, and William Robb MacLellan. 2019. “Single-Cell Imaging

and Transcriptomic Analyses of Endogenous Cardiomyocyte Dedifferentiation and Cycling." *Cell Discovery* 5 (June): 30.
Zheng, Grace X. Y., Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, et al. 2017. "Massively Parallel Digital Transcriptional Profiling of Single Cells." *Nature Communications* 8 (1): 14049.
Zhou, Jian, Chandra L. Theesfeld, Kevin Yao, Kathleen M. Chen, Aaron K. Wong, and Olga G. Troyanskaya. 2018. "Deep Learning Sequence-Based Ab Initio Prediction of Variant Effects on Expression and Disease Risk." *Nature Genetics* 50 (8): 1171–79.
Zhou, Xiang, Carolyn E. Cain, Marsha Myrthil, Noah Lewellen, Katelyn Michelini, Emily R. Davenport, Matthew Stephens, Jonathan K. Pritchard, and Yoav Gilad. 2014. "Epigenetic Modifications Are Associated with Inter-Species Gene Expression Variation in Primates." *Genome Biology* 15 (12): 547.
Zimmerman, Kip D., Mark A. Espeland, and Carl D. Langefeld. 2021. "A Practical Solution to Pseudoreplication Bias in Single-Cell Studies." *Nature Communications* 12 (1): 1–9.

VITA

David Read was born in Midland, Michigan and lived there until college. He participated in a variety of hobbies including Boy Scouts, theater, marching band, and playing tennis. His parents encouraged his interests in science from an early age and were a powerful voice for “lab” safety during backyard chemistry experiments. David went to the University of Michigan for undergraduate studies and received a major in biochemistry and a minor in computer science, as well as joining the university’s club Tae Kwon Do team. Unsated after 17 years of continuous schooling, David entered the PhD program in Genome Sciences at the University of Washington the following fall and has enjoyed both research and life outside of lab for the past five years in Seattle.