Expanding the scope and utility of single-cell genomic technologies

Andrew J. Hill

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

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Jay Shendure, Chair

Cole Trapnell, Chair

Christine Disteche

Program Authorized to Offer Degree:
Genome Sciences

University of Washington

Abstract

Expanding the scope and utility of single-cell genomic technologies

Andrew J. Hill

Co-Chairs of the Supervisory Committee:
Professor Jay Shendure
Genome Sciences

Assistant Professor Cole Trapnell
Genome Sciences

Technological advancements in single-cell genomic technologies have led to an exponential increase in number of cells from which biologists are able to measure various molecular profiles. This dramatic increase in scalability opens up the potential to leverage such technologies as the basis for a wide array of off-label applications. For example, the genetic screening field has typically been limited approaches that examine changes in relative abundance of genotypes represented in a given cell population before and after screening/selection as a proxy for phenotype. Molecular measurements such as RNA sequencing (RNA-seq) could allow a more generic and direct readout of molecular phenotypes, but performing bulk RNA-seq on many mutants in an arrayed format dramatically limits scalability. Single-cell RNA-seq, if modified to enable the readout of cell genotypes, would allow us to examine the molecular changes resulting from individual genotypes in a pooled format. In fact, single-cell technologies create the tremendous opportunity of multi-modal measurements (e.g. gene expression in addition to genotypes, lineage information, protein abundances via oligo tagged antibodies, T and B cell receptor sequences, entire additional molecular measurements like chromatin accessibility). An important step in the progression towards this goal is the adaptation of other molecular profiling techniques beyond RNA sequencing to single-cell formats, which come with their own experimental and analytical challenges that have yet to be

solved within the field.

In this dissertation, I first introduce an attempt to use single-cell transcriptomics to serve as a readout for genetic screens. In Chapter 2, I introduce our work to enable such a readout from CRISPR and CRISPRi-based genetic screens and find that a number of approaches that have been used to tackle this problem suffer from high error rates (50%) in the assignment of correct genotypes to cells. Ultimately we recommend a particular existing design that does not suffer from this common flaw. In Chapter 3, we utilize this method in conjunction with highly multiplexed perturbations to enable the screening of almost 6000 putative regulatory elements in the genome, measuring their impact on gene expression in *cis*. Notably an experiment of this scale would have been extremely difficult to achieve without the use of single-cell technologies. Lastly, in Chapter 4, I shift focus to our efforts to expand the set of molecular measurements that we can make with single-cell technologies. Specifically, I describe our efforts to measure chromatin accessibility using single-cell ATAC-seq across 13 different tissues in 8-week old mice and how we went about addressing the many computational challenges that arise when attempting to analyze and interpret such datasets.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Table Number Page

ACKNOWLEDGMENTS

I owe a great deal to the numerous people who supported me during the completion of this work. In graduate school there were often long, difficult periods where the science just wouldn't cooperate and it is only due to my advisors, colleagues, friends, and family (both human and otherwise) that I have been able to arrive at this point successfully. I want to emphasize that despite any perceived level of success or lack of adversity, the vast majority of (or perhaps all) graduate students face many difficult times. I was unlucky enough to struggle for a year and a half on my first main project only to be scooped five times over, and fortunate enough to have colleagues that helped support me as I got myself out of a substantial rut. Only with the help of the people around me did I go on to lead and contribute to the work in this dissertation that I am immensely proud of. Determination and persistence will only take you so far without a supportive lab environment, and I was incredibly lucky to have both.

First, I would like to thank my advisors Jay Shendure and Cole Trapnell. Jay, you have been incredibly generous in your advising and afforded me the uncommon opportunity to have a mentor who is both one of the nicest people I have ever met and flexible in allowing me pursue formative experiences outside the lab (e.g. internships and consulting) that have helped me tremendously in my own personal growth perhaps at the expense of some degree of short-term productivity in lab. Cole, you have been a constant advocate and and provided excellent career advice that has been invaluable as a young scientist. It has been incredible to watch you build such a successful research program from scratch. You have been extremely generous with your time and there were countless occasions when your hands on mentorship in analysis/biology was crucial to my own success and learning. Both of you have a seemingly bottomless capacity for finding silver linings and having a positive outlook in science. I don't know that I can fully put into words what a gift

this attribute has been. I tend to be very critical of my own work and progress so having you as constant cheerleaders in my life through both good times and bad has been invaluable.

I would also like to thank the other members of my supervisory committee, William Noble, Ray Monnat, David MacPherson, Patrick Paddison, and Christine Disteche for support, important scientific input, and career advice. Also, thank you to all my past scientific advisors: Daniel MacArthur, Monkol Lek, Blake Hannaford, and Howard Chizeck as well as Chris Neils and Kelli Jayne Nichols.

I have been incredibly fortunate to be part of the Shendure and Trapnell Labs, which are each home to amazing groups of smart, engaged, kind, and supportive scientists. Thank you so much to both labs in their entirety, although I want to thank (in no particular order) a few people specifically: Molly Gasperini, Charlie Lee, Matthew Snyder, Martin Kircher, José McFaline-Figueroa, Darren Cusanovich, Riza Daza, Delasa Aghamirzaie, Hannah Pliner, Fanny Huang, David Read, Silvia Domcke, Sam Regalado, Lea Starita, Jonathan Packer, Dana Jackson, Beth Martin, Seungsoo Kim, Melissa Zhang, Anh Leith, Xiaojie Qiu, Jun Cao, Sanjay Srivatsan, Greg Findlay, Aaron McKenna, Lauren Saunders, and Mike Morse. Also thank you to several folks in our department more broadly: Melissa Chiasson, Max Dougherty, John Lazar, and Kenneth Matreyek. I am extremely lucky to have had such a supportive group of colleagues and friends.

Genome Sciences has been a fantastic environment for graduate training. Thank you to Bob Waterston and the entire GS faculty for making it such a welcoming and supportive environment. Also, thank you to Francis Cheong, Bob Waterston, and David Hawkins for mentorship as a teaching assistant within the department.

In the summer of 2016 I organized an internship with 10X Genomics, which ended up being a transformative experience for me professionally. Thank you to Jay and Cole for supporting me in pursuing this work outside the lab. Also thank you to the amazing group of scientists I had the opportunity to work with during my time at 10X and specifically Deanna Church, Michael Schnall-Levin, Tarjei Mikkelsen, Ben Hindson, Patrick Marks, Grace Zheng, Paul Ryvkin, Jessica

Terry, Joe Shuga, Phil Belgrader, Solongo Ziraldo, Luz Montesclaros, Daniel Riordan, Sofia Kyriazopoulou Panagiotopoulou, Matthew Sooknah, Vijay Kumar, Alex Wong, David Stafford, David Lin, Kevin Wu, and Jeff Mellen.

Most people come to graduate school and might not know anyone or have anyone to share the ups and downs with. I am incredibly fortunate to have joined not only the same department, but also the same lab, as my partner, Molly. I'll never forget Jay's relief when we told him we were getting married. Molly, I can't possibly imagine having taken on this challenge without you by my side and I still can't believe how lucky we have been to share so many amazing experiences over the past few years. I'm so excited to see what we'll do together as we move on to yet another big challenge.

Thank you to my family, who have, much like Jay and Cole, been persistent cheerleaders and fans of Molly and I. Thanks Leslie, Roger, Ian, the many cats over the years (Zippidy, Nutmeg, Willow, Frisky, Ragamuffin, Twizzle, Rosie, and Amber) for all your support over the years and my loving extended family John, Betty, Liz, Dave, Charlie, Peggy, Pat, Sue, C.J., Jim, Bill, John O., Kelly, Christian, Lee, Owen, McKinley, Wesley, Abigail, Atticus, Allison, and Molly Kate.

While many people contributed to my happiness in graduate school, cats (and dogs too I guess) have also played an important role. Stella, Herman, Hubert, and Pippa, thank you for your bottomless enthusiasm for pets and emotional support. To be more specific, Stella is the best cat ever and tried tirelessly to prevent me from doing any actual work even though her implicit emotional support actually made me way more productive in the long run. Documentation of Stella's contributions to the work in this dissertation can be viewed here: `https://photos.app.goo.gl/wXUpssf2VsK2H2ji9`.

Chapter 1: INTRODUCTION

Since the initial adoption of massively parallel, often referred to as ”next generation”, sequencing in the early 2000's, a litany of different assays have been developed for converting biochemical and biophysical properties cells into digital assays that rely on counts of DNA molecules as their readout.

In large part, this change in paradigm has been motivated by the dramatically increased ability of these methods to make genome-wide measurements. For example, while chromatin immunoprecipitation (ChIP) followed by qPCR had been used as an analog readout of relative presence of a given epitope at an individual site or set of sites in the genome (e.g. transcription factor binding or the deposition of histone marks), the poor scalability of this class of techniques preclude their application genome-wide. While micro-array based approaches such as ChIP-chip provide potentially more high-throughput measurements (Buck and Lieb, 2004; Mockler and Ecker, 2005), they have many technical limitations and have largely been replaced by sequencing-based approaches. ChIP read out by sequencing (ChIP-seq) instead utilizes a direct readout of the DNA sequences of fragments enriched in the process of ChIP, ultimately mapping of reads to the genome, enabling detection of regions of the genome that have differential signal between samples (Albert et al., 2007; Barski et al., 2007; Mikkelsen et al., 2007; Johnson et al., 2007). In fact, the measurement of many different properties within cells have likewise been converted to digital, sequencing-based readouts (e.g. RNA sequencing, chromatin accessibility, proximity ligation, etc.).

The widespread adoption of these diverse techniques within both individual studies and large consortium efforts (ENCODE Project Consortium, 2012) has resulted enormous datasets consisting of many different assays performed across many different cell-lines, sorted cell populations, or bulk tissues in attempt to comprehensively characterize molecular state. While these datasets and techniques have been and will continue to be incredibly useful, there are several contexts in which their application is limited. First, in the case of bulk tissues or developing organisms, the measure

of bulk assays is an aggregate measure of the datatype over the diverse set of cell types present in the tissue, meaning that it is not representative of any one cell type in the sample. Second, in many applications one is examining the response of a given cell population throughout a dynamic process or stimulus, which is complicated by Simpon's paradox (Trapnell et al., 2014). Third, there are many applications in which one could imagine reading out auxiliary information associated with different subsets of cells in a population (say measures of lineage or genetic perturbation within a heterogeneous population), that are difficult or impossible to employ when making bulk measurements.

Single-cell genomic technologies, which use a variety of different techniques to measure molecular profiles of individual cells rather than bulk populations, have recently been introduced as an alternative that seeks to remedy these limitations. While there were many early challenges in field of single-cell genomics, perhaps one of the most important to the increased adoption of single-cell technologies has simply been achieving scalability – that is profiling a sufficiently large number of cells at reasonable cost.

1.1 Progression of scalability in single-cell RNA-seq technologies

Single-cell RNA-sequencing (scRNA-seq) was the first sequencing-based genomic technology to be converted to a single-cell format and the scalability of available scRNA-seq assays has grown exponentially over the past decade (Svensson et al., 2018). The earliest example of scRNA-seq was performed simply by preparing cDNA libraries from an individual cell in a tube (Tang et al., 2009), a study motivated by the paucity of methods that could handle low amounts of input material in the generation of cDNA libraries. However, the focus of the field towards the utility of making single-cell measurements would eventually shift as labs showed that single-cell qPCR data could be used to identify cell types (Guo et al., 2010), later followed by unbiased scRNA-seq (Islam et al., 2011). Over the next five years or so, in addition to many improvements in chemistry, the use of well plates, automated well plate protocols, and partially automated microfludic capture solutions lead to datasets in the of up to one thousand to ten thousand cells (the history of this progression

is covered in detail in Svensson et al. (2018)). However, protocols were still very laborious and/or costly, limiting their scalability and adoption beyond a relatively small number of labs.

In 2015, borrowing from techniques for digital droplet PCR (Vogelstein and Kinzler, 1999; Hindson et al., 2011), both Macosko et al. (2015) and Klein et al. (2015) showed that cells rather than DNA molecules could be co-captured along with a oligonucleotide labeled beads in an emulsion of aqueous phase droplets partitioned by oil. In these methods, each droplet serves as a reaction chamber encapsulating the nucleic acids of each cell along with reagents required for reverse transcription. RNA molecules from each cell are reverse transcribed using a barcoded oligo-DT RT primer that is released from the bead, effectively tagging each resulting cDNA molecule with the barcode associated with each bead. This enables the measurement of scRNA-seq profiles from a pooled library prep. A generic diagram of droplet-based methods is included in Figure 1.1. A necessary limitation of these methods is that because all sequenced fragments must also be tagged with this cell barcode, libraries prepared using these methods only include the three-prime ends of cDNA molecules (a property shared by all current high-throughput RNA-seq methods). These techniques now exist in several commercial forms, with 10X Genomics (Zheng et al., 2017) one of the most widely utilized to date. The combination of improved scalability, reduced costs, and ease of use have lead to the widespread adoption of droplet-based single-cell RNA-seq methods that are now capable of generating datasets as large as one million cells (datasets in the 10,000-100,000 range are now fairly routine in the literature). As of this writing, a lack of serious competition in this space has led to high commercial costs of reagents. As more competitors enter this space and alternative protocols continue to emerge, we can expect that reagent costs will likely drop, enabling larger scale experiments to become more tractable.

More recently, we have also seen the emergence of several other paradigms as alternatives to droplet-based technologies. For example, several studies have utilized co-capture of cells and oligonucleotide labeled beads in nano/micro-wells on a chip as a way of increasing throughput relative to well plate-based technologies (Gierahn et al., 2017; Han et al., 2018). Microwell technologies are in the early stages of commercialization.

Our lab has also developed a completely different paradigm commonly referred to as combina-

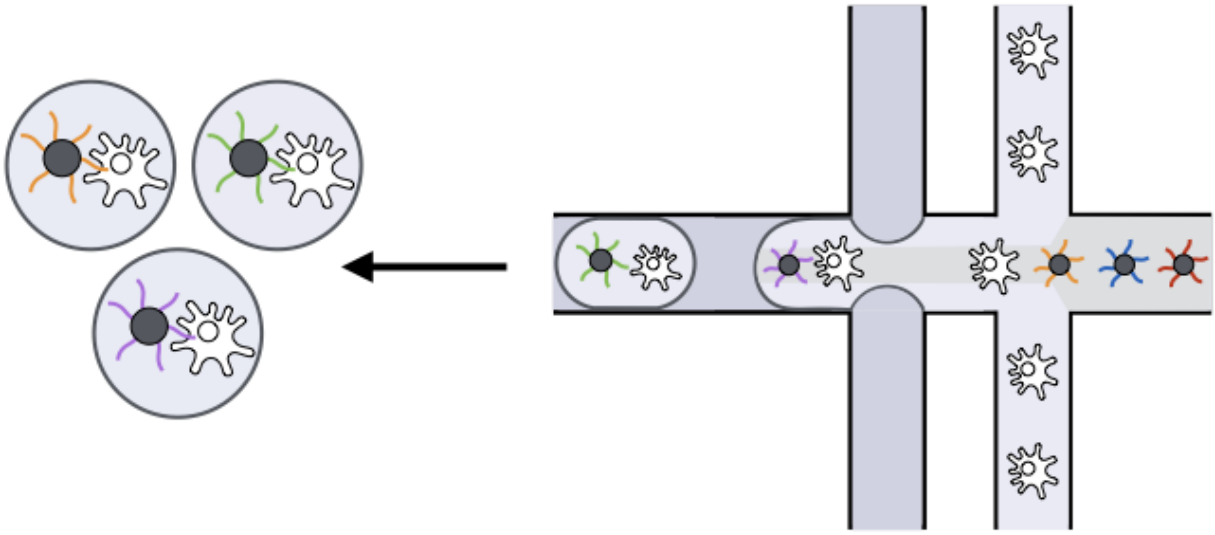


Figure 1.1: Diagram of droplet platforms for single-cell genomics. Detailed descriptions of different implementations of droplet-based platforms are available in Macosko et al. (2015), Klein et al. (2015) and Zheng et al. (2017). Briefly cells are combined with oligonucleotide labeled hydrogel beads that harbor a DNA sequence unique to each bead. In Macosko et al. (2015) and Klein et al. (2015), both cells and beads are loaded into droplets at a concentration-dependent rate according to poisson statistics. Zheng et al. (2017) improves upon this method by ensuring that every droplet receives a bead and only cells are loaded at a rate determined by their loading concentration, greatly increasing the overall cell capture rate and reducing the duration of capture. The microfluidic device also typically generates an emulsion in oil such that an aqueous-phase partition is generated around each cell. Hydrogel beads harbor a barcoded DNA molecule that also facilitates a given step of the biochemical reaction (in the case of single-cell RNA-seq this is typically a barcoded oligo-DT RT primer). Once the relevant nucleic acids are barcoded (cDNA molecules in single-cell RNA-seq), the emulsion can be broken and libraries can be processed in a pooled fashion. Several advancements have also been made in this space to enable more complex multi-step protocols.

torial indexing or split-pool (Amini et al., 2014; Cusanovich et al., 2015), which was later adapted to scRNA-seq (Cao et al., 2017). These methods first deposit many cells per well into an initial 96 or 384 well plate and receive an initial round of in-situ barcoding specific to each well such as a barcode oligo-DT primed RT step in the case of single-cell RNA-seq. Since the cells are still intact, they can be repooled and split back out into a second set of wells for further rounds of barcoding, such as indexed PCR or barcoded ligation steps. During this process, while many cells may share a given round of barcoding, the randomization of cell groupings between barcoding rounds means that with proper titration of the number of cells in each well, most cells will have traveled through a unique combination of wells and thus harbor a unique combination of barcodes. A general diagram of sci-* methods is provided in Figure 1.2. These methods have the benefit of increased flexibility (beads are replaced by simple DNA oligos) and with 3 or more rounds of indexing have been shown to be highly scalable. Currently the main tradeoff of this class of methods is decreased library complexity and laborious and difficult protocols, but efforts are underway to make these protocols more widely exportable.

A figure illustrating the growth of scRNA-seq method scalability (using the size of published datasets as a proxy for feasibility of experiments) is included as Figure 1.3 (adapted from Figure 1 of Svensson et al. (2018)). As the set of usable technologies continue to expand and the cost of sequencing continues to drop (note that sequencing often accounts for half or more of total costs in many cases even with current pricing of commercial platforms depending on the assay), the scale of experiments that we can feasibly achieve with single-cell genomic technologies opens up tremendous new possibilities.

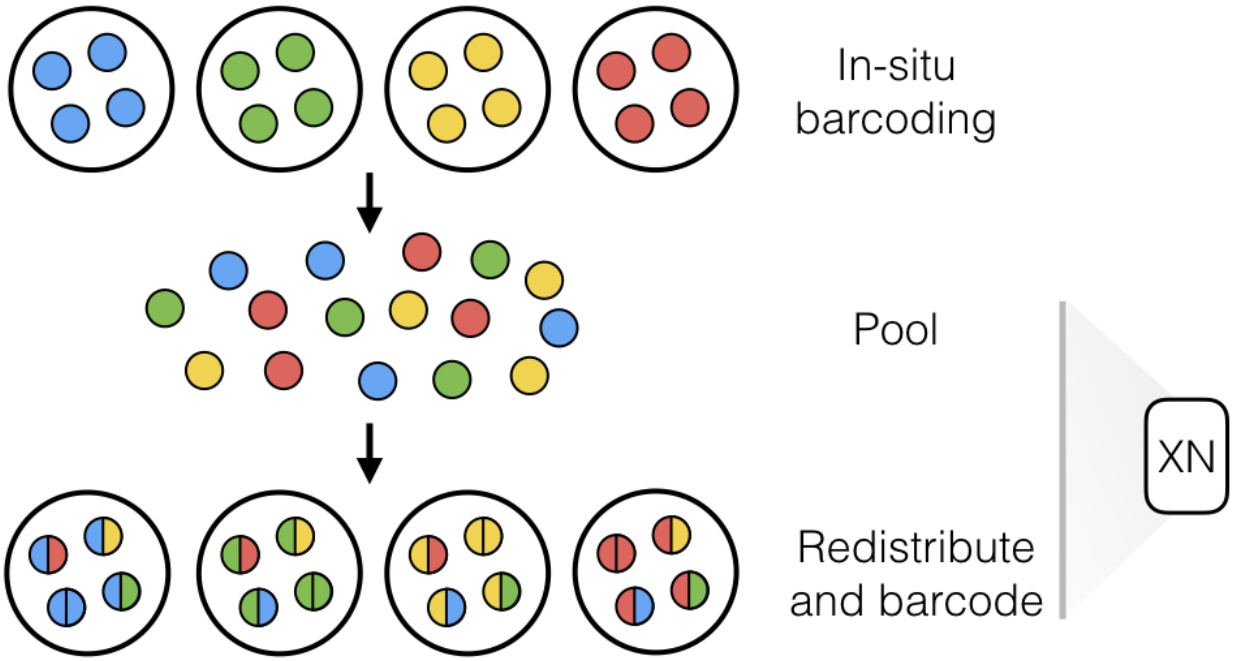


Figure 1.2: Generic diagram of sci-* methods for single-cell genomics. Cells are typically plated in well plates and undergo a round of *in-situ*, such that a single barcode is added to each DNA or RNA molecule in the cell. The cells are then pooled back together and intermixed before being redistributed into a second set of wells for further barcoding. Note that the final round of barcoding need not be *in-situ* as at this point no further split pool rounds are required after this point. Note that while many cells will share a the same barcode sequence in a given round of barcoding, the combination over all rounds will be unique for most cells as long as the number of cells per well in each round is carefully calibrated according to the expected collision rate given the total number of possible unique barcode combinations. Both the total number of wells and number of barcoding rounds can be used to modulate the scalability of the assay. Several different single-cell assays have been implemented using this general paradigm.

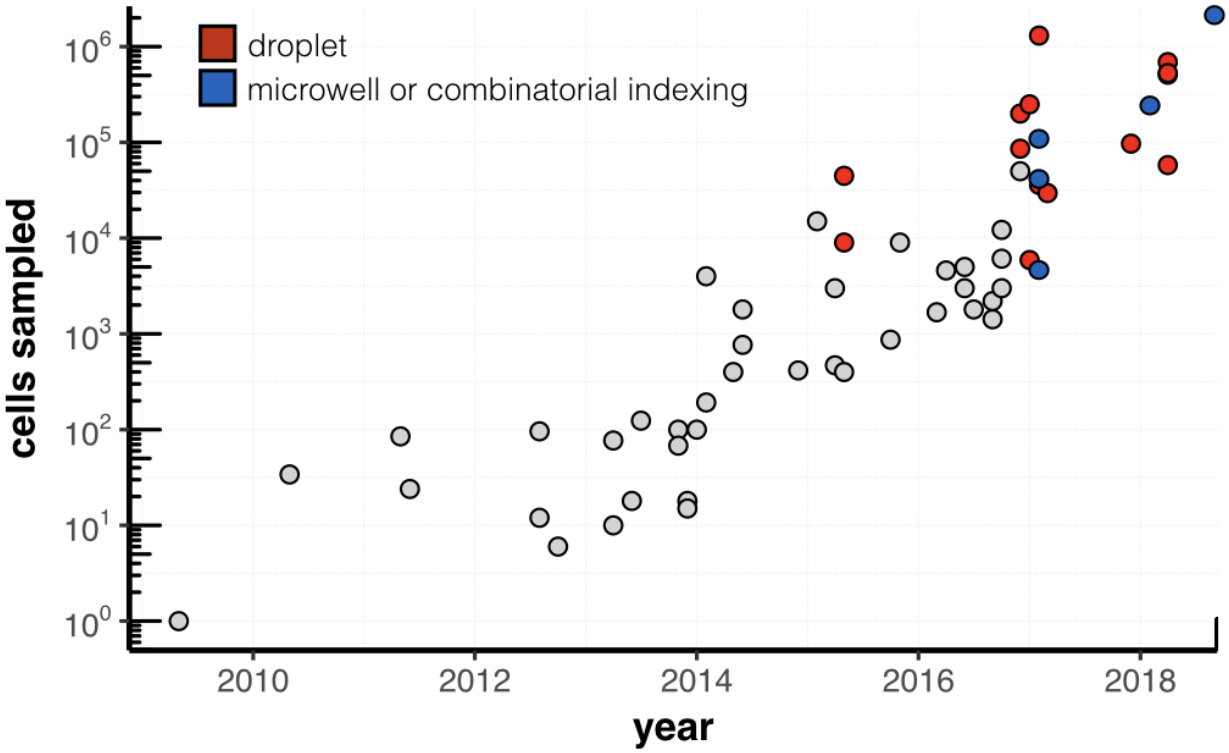


Figure 1.3: Exponential scaling of scRNA-seq methods in the past decade. Figure adapted from Figure 1 of Svensson et al. (2018). A number of representative single-cell RNA-seq studies published from 2009 to 2019 are shown, plotting the number of cells in the published dataset and the year of the publication. In this case, number of cells reported is used as a proxy for what was feasible around the time of publication (in many cases likely with a year or more lag from data collection to publication). Importantly, these experiments are not a random subset and have been chosen for their representative scalability. Usage of droplet-based platforms (red) in the literature as of this writing vastly outweighs other high-throughput paradigms (blue) in contrast to their approximately equal representation in this plot (although as with any area of technology this may continue to change over time).

1.2 UTILIZING SCALABLE SINGLE-CELL PARADIGMS

While there are some fairly obvious uses of single-cell technologies, such as profiling relative transcript abundances to catalog cell types across the many tissues of humans and model organisms, there are a number of novel approaches and experimental designs that could be enabled by single-cell measurements.

1.2.1 Potential for the application of single-cell genomics to genetic screens

While next generation sequencing has been used to generate measurements of biochemical and biophysical properties within a cell, ultimately massively parallel sequencers can act as a generic counting device ina number of different contexts. One very important application has been pooled genetic screens for high-throughput genotype to phenotype experiments.

Briefly, this class of experiments, which are now numerous in their approaches and specific instances (Gasperini et al., 2016; Shalem et al., 2015; Mohr et al., 2014), introduce reporters, mutant proteins, genetic variation in the genome, etc. into cells (mammalian cells are common, experiments are also commonly done in yeast and other model organisms). Some subset of these cells will take on an altered phenotype in the appropriate context. Because the ultimate readout of these assays is relative abundance of variant representation in the population (or sections of the population as determined by a reporter), the experiment must be able to use screening or selection as a proxy for the underlying phenotype. Many such assays are either generic but limited to a small number of genes (e.g. essentiality or growth phenotypes) or targeted very specifically to their pathway or protein of interest. An example of this general paradigm is illustrated in Figure 1.4.

Ultimately, more generic readouts would greatly benefit the genetic screening community. More recently, work has been done to develop more generic assays based on protein stability (Matreyek et al., 2018), although this approach cannot measure the effects of mutations that impact protein function but not stability.

Given that the altered phenotype typically measured in genetic screens would often originate

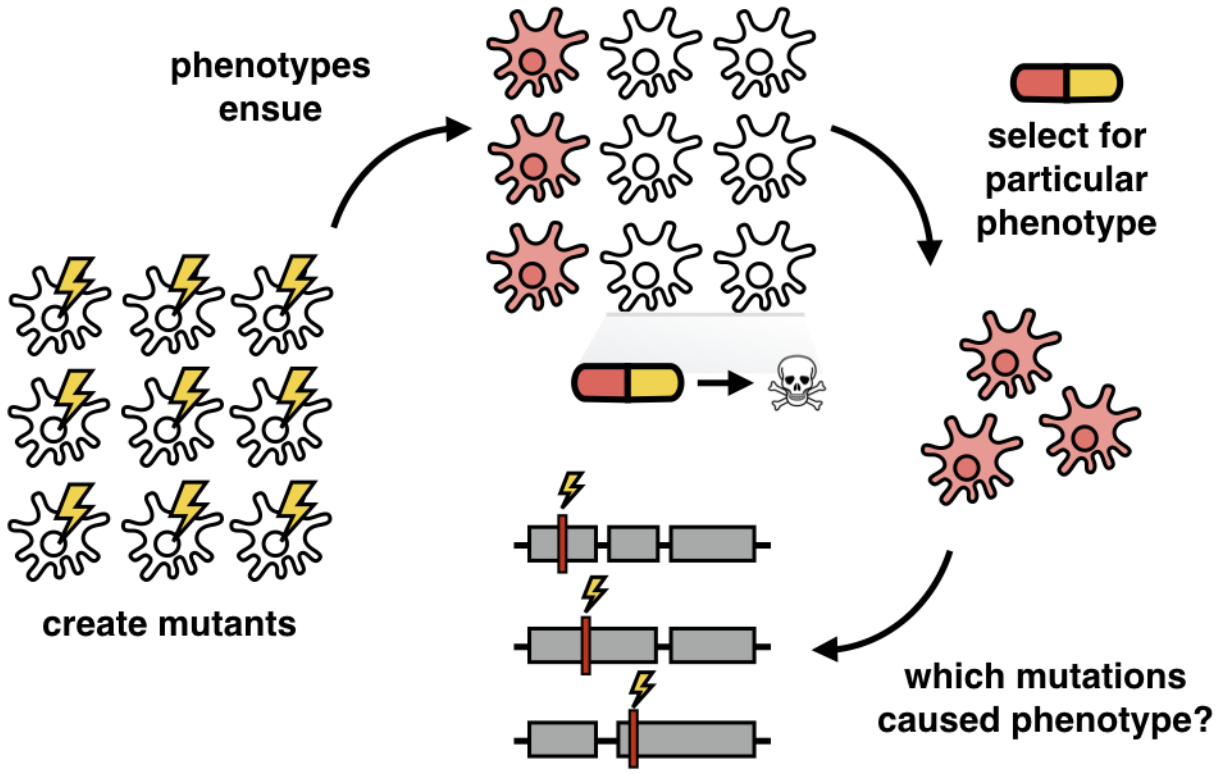


Figure 1.4: Pooled genetic functional assays typically rely on screening or selection. High-throughput genetic screening methods often introduce one or more mutations into each cell in a population of cells (often, but not always, *in-vitro*). Under a given set of conditions, a subset of genotypes will result in an altered phenotype. An assay to select or screen cells in a population for this altered phenotype has to be designed. A very simple example would be drug selection (shown in figure) or essentiality. For example, perhaps only cells with a mutation conferring drug resistance will survive a drug treatment. The differences in genotype representation before and after selection can be used to infer genotypes that gave rise to drug resistance via direct genotyping (sequencing coding sequence of gene being tested) or sequencing of barcodes in cases where a set of barcodes have been linked to mutations.

from or result in an altered molecular phenotype (e.g. changes in transcription), one alternate approach could be the use of genomics assays like RNA-seq to measure the impact of genetic mutations. However, performing bulk RNA-seq on many mutants individually would severely limit throughout by necessitating an arrayed rather than pooled format. Ideally, we would be able to read out RNA-seq (or other molecular assays) for individual mutations in the context of a pooled screen. In theory this should be possible using scRNA-seq, so long as we are able to obtain a proxy for each cell's genotype in addition to a transcriptional profile and therefore sort cells by genotype *in-silico*.

In addition to potentially more abstract molecular phenotypes (e.g. inferring what has happened via up/down-regulated pathways and genes), there are also screens where directly measuring changes in transcription for a specific set of nearby genes is the primary goal. For example, many screens in recent years have sought to connect enhancers/regulatory elements to the genes they regulate (Canver et al., 2015; Diao et al., 2016, 2017; Fulco et al., 2016; Gasperini et al., 2017; Klann et al., 2017; Korkmaz et al., 2016; Rajagopal et al., 2016; Sanjana et al., 2016). All of this prior work has been focused on a very small number of loci because specialized assays or GFP tagging was used to enable indirect measurement of gene expression via selection or screening within a pooled screen. scRNA-seq could potentially allow for direct readout of changes in gene expression genome-wide for any mutant.

1.2.2 Expanding the scope of single-cell technologies to other modalities

Although measurements of RNA have been extremely useful and popular (and will remain so for some time) in both bulk and single-cell studies, there are other measures that have been shown to be necessary to more completely characterize the molecular state of a cell. For example, while scRNA-seq can tell you the relative abundances of various RNA molecules in a cell, it tells you very little about the regulatory process that lead to transcription in the first place. Many bulk technologies have been developed and applied at scale to examine quantities like chromatin accessibility, methylation status, transcription factor binding, and histone mark deposition in addition

to RNA abundances (ENCODE Project Consortium, 2012). Subsequent multi-modal studies have proved extremely useful and it is likely that some subset of these measurements would also prove useful in many single-cell experiments.

One of the first methods beyond RNA-seq to be adapted to a single-cell format was ATAC-seq (Buenrostro et al., 2013). While mamallian promoters determine the specificity of transcription in some cases, often this specificity is determined by distal regulatory elements, which scRNA-seq fails to measure directly. One proxy for several different categories of regulatory element activity is chromatin accessibility. Much of your genome is packaged tightly in closed chromatin, which is physically inaccessbile to cellular machinery. However, regulatory elements such as promoters and regions bound by transcription factors, as a prerequisite for their endogenous function, must be physically accessible to the macromolecules of the cell. Chromatin accessibility has typically been measured genome-wide by DNase-seq (Hesselberth et al., 2009). However, DNase-seq is technically challenging and has high input requirements that have limited its applicability to single-cell methods. An alternate method, ATAC-seq, utilizes the Tn5 transposon (Adey et al., 2010), which preferentially inserts loaded DNA sequences into open chromatin, allowing for the generation of sequencing libraries that are enriched for regions of high accessibility (ATAC-seq is illustrated in Figure 1.5).

ATAC-seq, in contrast to DNase-seq, is a very simple protocol and can be used on a small number of input cells (or even individual cells). This lead to the development of single-cell ATAC-seq (scATAC-seq) methods using on-chip microfludic capture (Buenrostro et al., 2015) and combinatorial indexing (Cusanovich et al., 2015) (sci-ATAC-seq). sci-ATAC-seq utilizes combinatorial indexing (as described in Figure 1.2) by using transposons that will insert DNA sequences containing both priming sequences and well-specific barcodes as a first round of *in-situ* barcoding, followed by a pooling step and subsequent PCR reaction with well-specific primers as the second round of barcoding (see Figure 1.6). This technique allows for relatively high-throughput measurement of single-cell chromatin accessibility profiles.

However, despite our ability to measure chromatin accessibility in single-cells the analysis of this datatype is much less well characterized than RNA-seq for the purposes of cell-type iden-

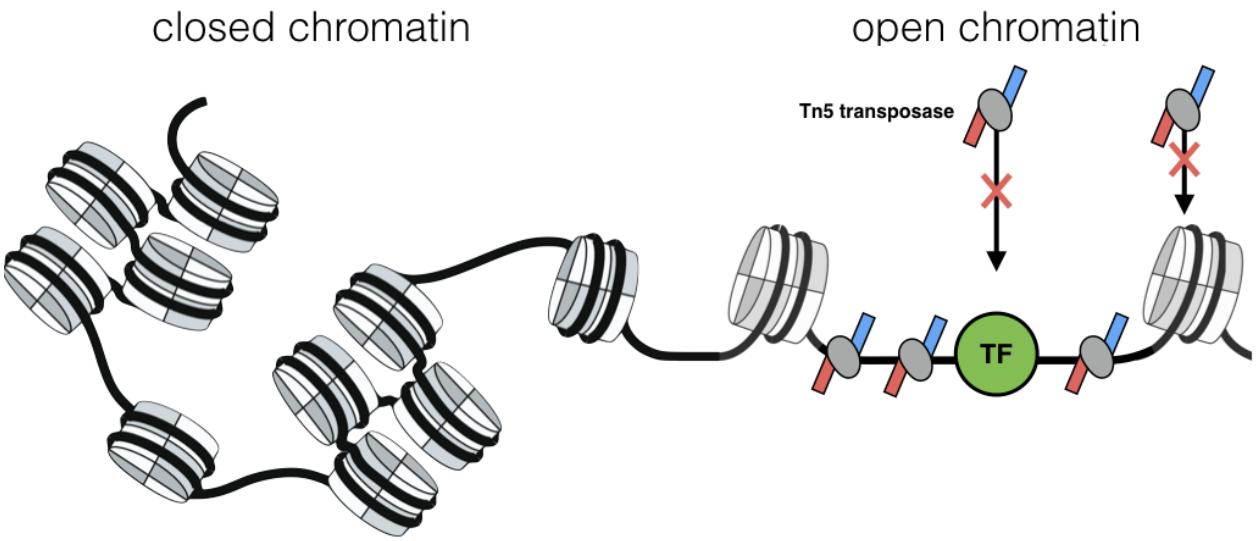


Figure 1.5: Diagram of ATAC-seq. The Tn5 transposon preferentially inserts loaded DNA sequences into regions of open chromatin. This protocol allows for the rapid generation of data very similar to DNase-seq but with lower input and labor requirements.

tification. There remain many challenges to both the optimization of more robust and scalable scATAC-seq protocols and the analysis of scATAC-seq data (e.g. sparsity and interpretability), both of which I have worked to address during graduate school.

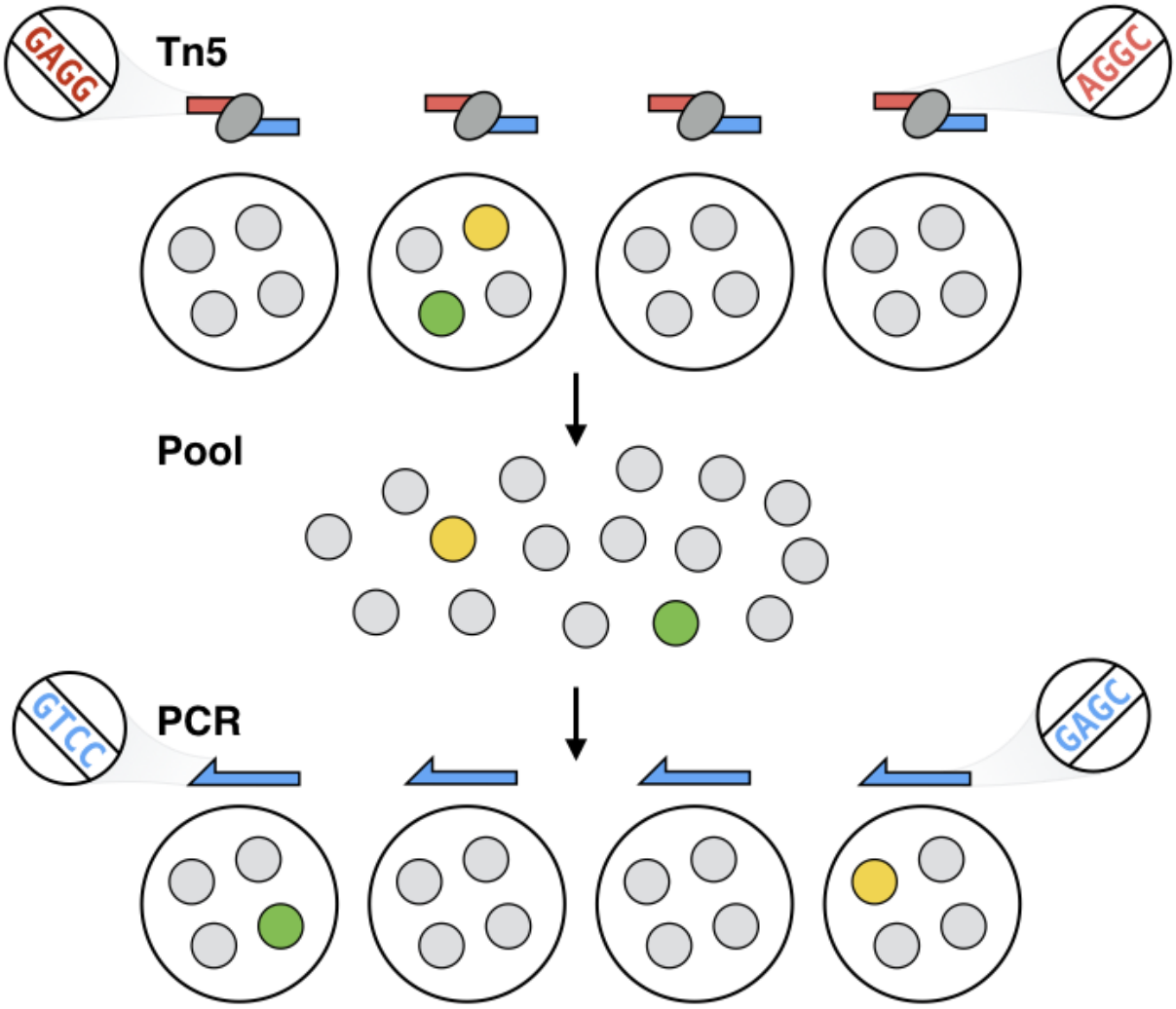


Figure 1.6: Diagram of sci-ATAC-seq. To adapt ATAC-seq to a combinatorial indexing format, Cusanovich et al. (2015) modified the protocol to utilize Tn5 transposons that harbored well-specific barcodes, carrying out the transposition reaction split across several wells with different loaded barcode sequences. Cells are pooled before being split out into a new set of wells for PCR with well-specific barcoded primers to serve as a second round of indexing. Note that the two cells labeled in green and yellow share the same starting well, but after two rounds of barcoding will have traveled through a unique combination of wells (and therefore have a unique combination of barcodes).

1.3 ORGANIZATION OF THIS DISSERTATION

Chapters 2 and 3 describe efforts to apply droplet-based scRNA-seq to read out molecular phenotypes in pooled genetic screens where otherwise homogenous cell populations are diverse with respect to their engineered genotype. Chapter 2 is focused on the technical aspects of enabling genotype readouts in scRNA-seq libraries. Chapter 3 describes the application of methods described in Chapter 2 to study the *cis* effects of expression when perturbing regulatory elements in the genome. Chapter 3 additionally introduces the concept of highly multiplexed random perturbations to measure the independent effect of several perturbations per cell. This dramatically increases the scalability of these screens, allowing us to measure the effects of almost 6000 different putative regulatory elements, many more than any past study.

In Chapter 4, I shift focus towards efforts to expand the set of molecular measurements that we can make with single-cell technologies. Specifically, I describe our efforts to measure chromatin accessibility using scATAC-seq across 13 different tissues in 8-week old mice and how I went about addressing the many computational challenges that arise when attempting to analyze and interpret such datasets.

Closing remarks and future directions are presented in Chapter 5.

Chapter 2: ON THE DESIGN OF CRISPR-BASED SINGLE-CELL MOLECULAR SCREENS

Chapter 2 is adapted with minimal modification from:

Hill, A.J.*, McFaline-Figueroa, J.L.*, Starita, L.M., Gasperini, M.J., Matreyek, K.A., Packer, J., Jackson, D., Shendure, J., Trapnell, C. (2018) On the design of CRISPR-based single-cell molecular screens. Nature Methods. *15*, 271–274.

2.1 ABSTRACT

Several groups recently coupled CRISPR perturbations and single-cell RNA-seq for pooled genetic screens. We demonstrate that vector designs of these studies are susceptible to ˜50% swapping of guide RNA–barcode associations because of lentiviral template switching. We optimized a published alternative, CROP-seq, in which the guide RNA also serves as the barcode, and here confirm that this strategy performs robustly and doubled the rate at which guides are assigned to cells to 94%.

2.2 MAIN

Pooled genetic screens based on RNAi or CRISPR enable thousands of programmed perturbations per experiment (Shalem et al., 2015; Mohr et al., 2014). However, assays for such screens are limited to coarse phenotypes (e.g., cell viability) and are uninformative with respect to the mechanism by which perturbations mediate their effects.

To circumvent these limitations, several groups recently reported using single-cell RNA-seq (scRNA-seq) as a readout for CRISPR-based pooled genetic screens. The single guide RNA (sgRNA) in each cell is identified together with its transcriptome, either via a Pol II transcribed barcode (CRISP-seq, Perturb-seq, Mosaic-seq (Xie et al., 2017; Adamson et al., 2016; Dixit et al., 2016; Jaitin et al., 2016)) (Figure 2.1A) or by capturing the sgRNA itself within a Pol II tran-

script (CROP-seq7) (Figure 2.1B). Toward similar goals, we pursued a lentiviral strategy similar to former methods (Xie et al., 2017; Adamson et al., 2016; Dixit et al., 2016; Jaitin et al., 2016) in which each sgRNA was linked to a barcode located several kilobases away (Figure 2.1A). In our vector (pLGB-scKO), the barcode was positioned in the 3' UTR of a blasticidin resistance transgene, enabling its recovery by scRNA-seq methods that capture poly(A) transcripts (Figure 2.2A,B). Guides and barcodes were paired during DNA synthesis, which facilitated pooled cloning and lentiviral delivery (Figure 2.2C).

With this design, we sought to ask how loss-of-function (LoF) of tumor suppressors altered gene expression in immortalized, nontransformed breast epithelial cells. We targeted TP53 and other tumor suppressors in MCF10A cells, with or without exposure to the DNA-damaging agent doxorubicin. Cloning and lentiviral packaging was performed either individually for each targeted gene ('arrayed') or in a pooled fashion. In addition to scRNA-seq, we performed targeted amplification (Dixit et al., 2016; Adamson et al., 2016) to more efficiently recover the barcodes present in each cell (2.2B and Figure 2.3).

With arrayed lentiviral production, a substantial proportion of cells in which TP53 was targeted had a gene expression signature consistent with failure to activate a cell cycle checkpoint response after DNA damage (e.g., lower expression of CDKN1A and TP53I3; Figure 2.4A). However, these effects were greatly reduced when we performed a similar experiment with pooled lentiviral production (Figure 2.4B). Furthermore, markedly fewer genes were differentially expressed in the pooled than in the arrayed experiment (Figure 2.4C). t-SNE embedding revealed that both experiments contained a cluster of cells characterized by expression of the mitotic marker CCNB2 and low levels of TP53I3, consistent with a TP53-null phenotype. In the arrayed experiment, this cluster was almost entirely composed of cells with sgRNAs targeting TP53 (99.4%). However, in the pooled experiment, only 41% of assigned cells from the corresponding cluster contained TP53 sgRNAs (Figure 2.4D-I).

We reasoned that lentiviral template switching may explain this difference. Lentiviral virions are pseudodiploid; i.e., two viral transcripts are copackaged during their production (Nikolaitchik et al., 2013; Tseng et al., 1997). The reverse transcriptase that acts before integration has a rate of

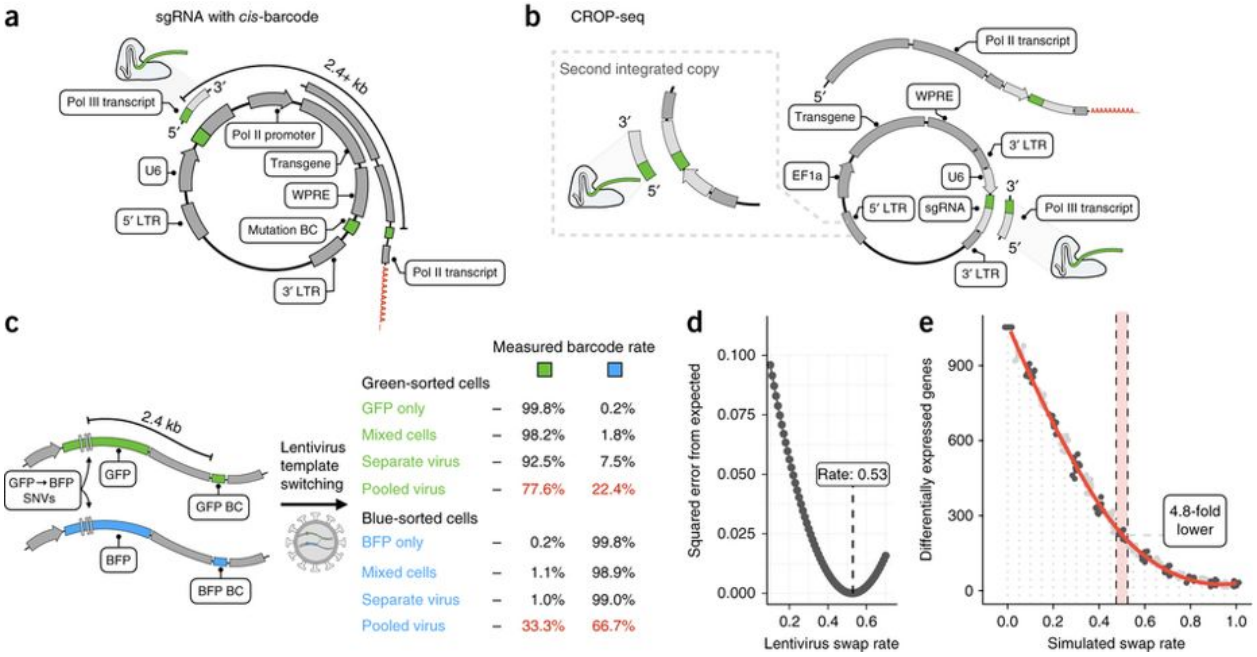


Figure 2.1: Template switching decreases the sensitivity of CRISPR-based single-cell molecular screens that employ linked barcodes. A) Schematic of vectors that rely on cis-pairing of sgRNAs and barcodes such as Peturb-seq, CRISP-seq, and MOSAIC-seq. A barcode (BC), expressed as part of the Pol II transcript and sequenced as a proxy for the guide sequence, is linked to an sgRNA by a distance of 2.4 kb or more. WPRE, woodchuck hepatitis virus post-transcriptional regulatory element. U6, a Pol III promoter. B) CROP-seq approach. One copy of the guide is cloned into the 3' LTR and transcribed as part of the Pol II transcript, which can be sequenced directly. A second copy of the guide expression cassette is produced in the 5' LTR during lentivirus positive-strand synthesis before integration. C) Template switching at 2.4 kb separation between the distinguishing bases (3-bp differences) in GFP and BFP and their respective barcodes. Percentages reflect sorted cells transduced with GFP virus (GFP only) or BFP virus (BFP only); these cells mixed before sorting (mixed cells); or cells transduced with mixed virus generated from GFP and BFP plasmid packaged individually (separate virus) or together (pooled virus). Note that in a mix of two plasmids, only approximately half of all chimeric products are detectable because of homozygous virions (see Online Methods). (*legend continued on next page*)

(*continued*) D) Sum of squared errors of observed data vs. expected values at various swap rates using the fraction of barcodes in the green and blue sorted samples (n = 4 measurements), assuming a relative proportion of 61.7% GFP+ cells as determined from FACS (see Figure 2.5 and Methods for details). E) Simulation of progressively higher fractions of target assignment swapping on data from the transcription factor pilot arrayed screen of Adamson et al. (2016), used here as a gold standard performed with arrayed lentivirus production. Number of DEG across the target label at FDR of 5% is plotted at each swap rate for ten samplings per swap rate (n = 5,321 cells used in tests). 0.5 corresponds to the 50% swap rate determined via FACS.

template switching (Jetzt et al., 2000) estimated as 1 event per kilobase (kb) (Schlub et al., 2010). In pooled lentiviral production, template switching should result in the integration of chimeric products at a rate proportional to the distance between paired sequences (Figure 2.5). This risk was noted by Adamson et al. (2016) and (Dixit et al., 2016). It was altogether avoided by (Adamson et al., 2016) through arrayed lentiviral production, but pooled lentiviral production was performed in some or all experiments of the other reports (Xie et al., 2017; Dixit et al., 2016; Jaitin et al., 2016). Although Sack et al. (2016) recently quantified this phenomenon at distances up to 720 bp in vectors designed for bulk selection screens, the implications of template switching at longer distances (e.g., the 2.5 kb+ separation between sgRNAs and barcodes in the pLGB-scKO, CRISP-seq, Perturb-seq, and Mosaic-seq vectors (Xie et al., 2017; Adamson et al., 2016; Dixit et al., 2016; Jaitin et al., 2016)), as well as for scRNA-seq study designs specifically, remain unexplored.

To test this hypothesis, we cloned BFP and GFP transgenes, which differ by 3 bp, into separate lentiviral vectors, pairing each with a unique barcode separated from the nearest unique bases in BFP or GFP by 2.4 kb (Figure 2.1C). We transduced MCF10A cells with lentivirus generated either individually or as a pool of the two plasmids, FACS-sorted GFP+ or BFP+ fractions, and we quantified the rate of barcode swapping (Figure 2.1C and 2.6). At this distance, swapping occurred at the theoretical maximum rate of 50% (Figure 2.1D and Figure 2.7).

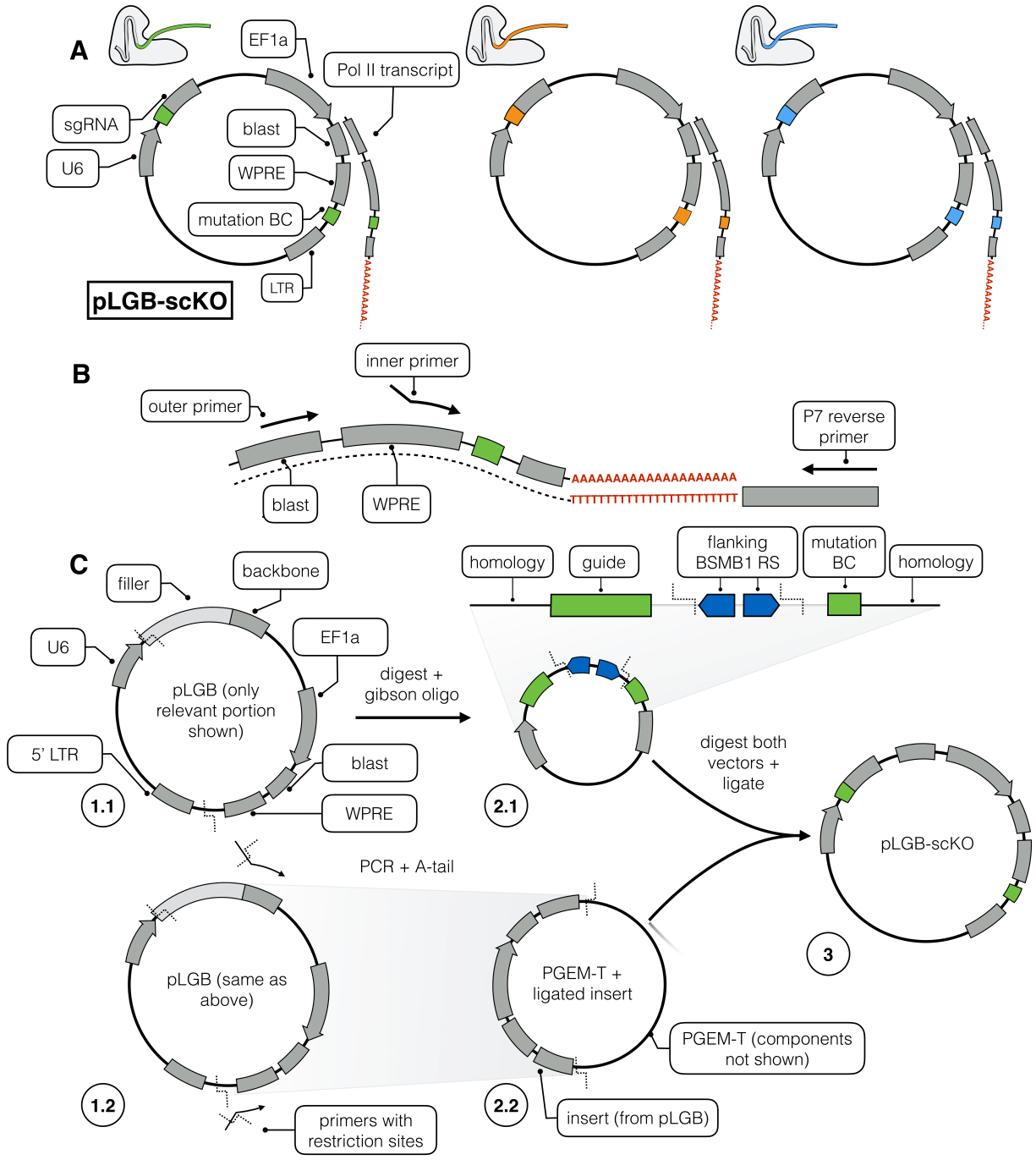


Figure 2.2: Diagram of cloning protocol and barcoded transcript enrichment strategy relying on cis pairing of sgRNAs and barcodes (pLGB-scKO). Details on exact restriction sites and primer sequences used can be found in the methods. A) Schematic of our final vector relying on cis pairing of an sgRNA and a distal barcode. B) Nested PCR enrichment strategy to maximize recovery of barcoded transcripts from single- cell RNA-seq data. (*legend continued on next page*)

(*continued*) C) Pooled cloning protocol. In 1.1 we start with pLentiguideBlast and digest near the final locations of the sgRNA and paired barcode. In 2.1 an engineered library of oligos containing programmed pairs of sgRNAs and corresponding barcodes are inserted into the digested vector. In 1.2 a portion of pLentiguideBlast is amplified. In 2.2 this fragment is cloned into PGEM-T. Finally, in step 3 vectors resulting from 2.1 and 2.2 are digested with Bsmb1 and the insert from 2.2 is ligated into the backbone in 2.1 to produce the final library of sgRNAs and paired barcodes.

To simulate the impact of template switching, we obtained data from Adamson et al. (2016) generated using the Perturb-seq vector with arrayed lentiviral production. We swapped target labels in silico at varying rates, and we evaluated power to detect differentially expressed genes (DEG). With 50% swapping, we observe a 4.8-fold decrease in the number of DEG (Figure 2.1E). This loss in power results from an effective two-fold reduction in number of useful cells per target, coupled with noise from swapped associations.

CROP-seq (Datlinger et al., 2017) differs from the other methods (Xie et al., 2017; Adamson et al., 2016; Dixit et al., 2016; Jaitin et al., 2016) in that it does not rely on pairing of sgRNAs and barcodes. Instead, the sgRNA itself serves as a barcode as part of an overlapping Pol II transcript. Furthermore, the sgRNA cassette is copied from the 3' to 5' long terminal repeat (LTR) during positive-strand synthesis (Figure 2.1B) via an intramolecular priming step that does not result in appreciable intermolecular swapping (Yu et al., 1998). A limitation of CROP-seq is that sgRNAs are recovered from scRNA-seq data with limited sensitivity (˜40–60%)7, such that half the single-cell transcriptomes are discarded. We modified CROP-seq to include targeted amplification of the sgRNA region from mRNA libraries already tagged with cellular barcodes, similar to our pLGB-scKO design (Figure 2.8A,B).

To evaluate this approach, we performed a CRISPR-mediated LoF screen of 32 tumor suppressors (six guides per target) and six nontargeting control (NTC) guides in MCF10A cells with or without doxorubicin. Whereas sgRNA(s) would generally be identified at a rate of 42–47%

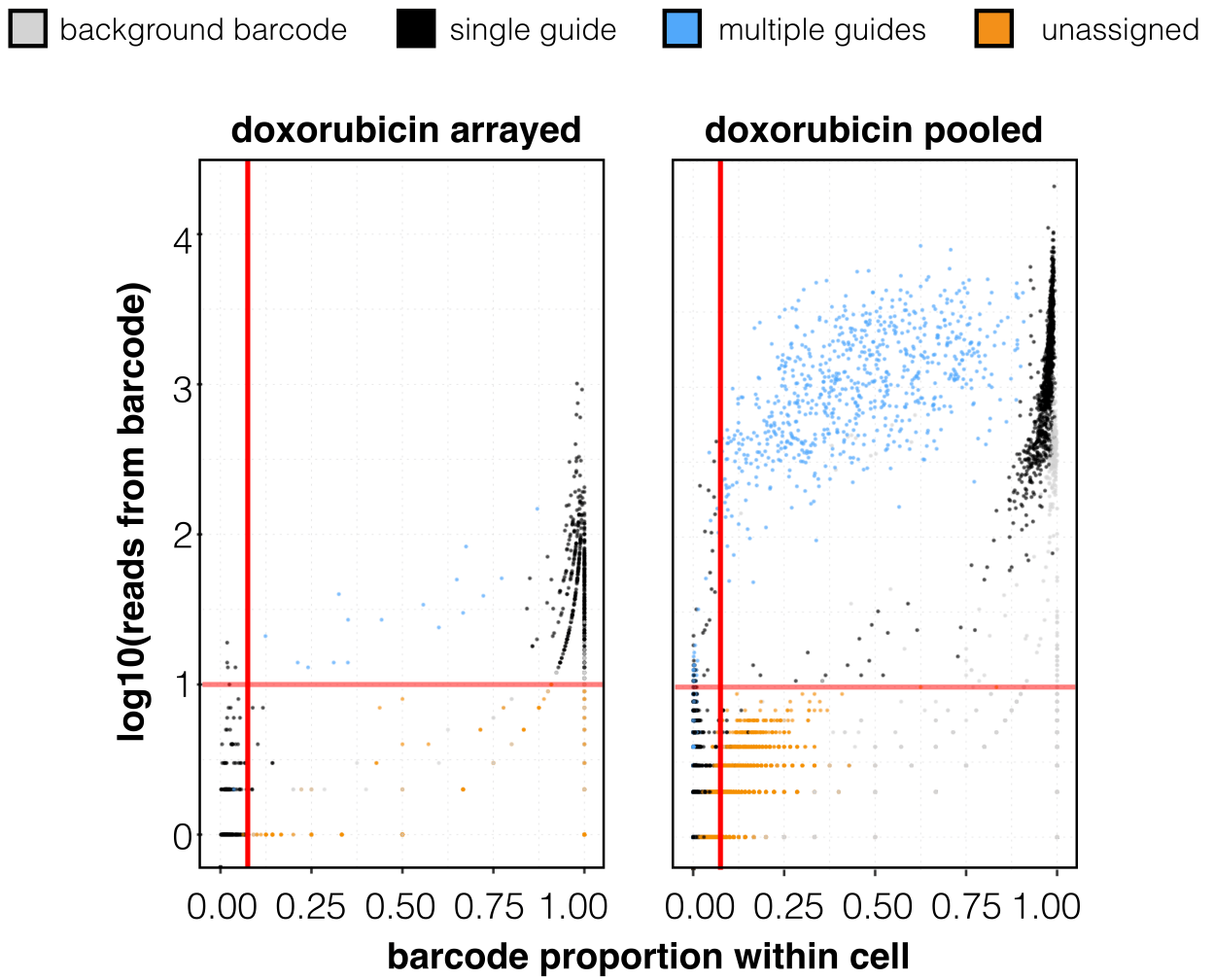


Figure 2.3: Barcoded transcript enrichment quality control for arrayed and pooled pLGB-scKO experiments. Each dot represents a barcode sequence observed in a given cell. Plot of reads for a given barcode against the proportion of all barcode reads observed in a given cell for every barcode/cell pair colored according to whether the cell barcode is determined to be background from whole transcriptome scRNA-seq data, a cell with a single guide assignment, a cell with multiple guide assignments, or a cell that ultimately receives no assignment given the applied thresholds. Red lines indicate the lower-bound thresholds used to distinguish noise from true barcode observations (10 reads and 0.075 proportion within cell). All barcodes observed above the red lines are assigned to their respective cells. Left, doxorubicin treated sample from arrayed experiment. Right, Doxorubicin treated sample from pooled experiment. Note that the rate of cells with single guide assignments in the doxorubicin treated sample is approximately 81%, which is in line with the 80% expected value at an MOI of 0.45 in a selected population.

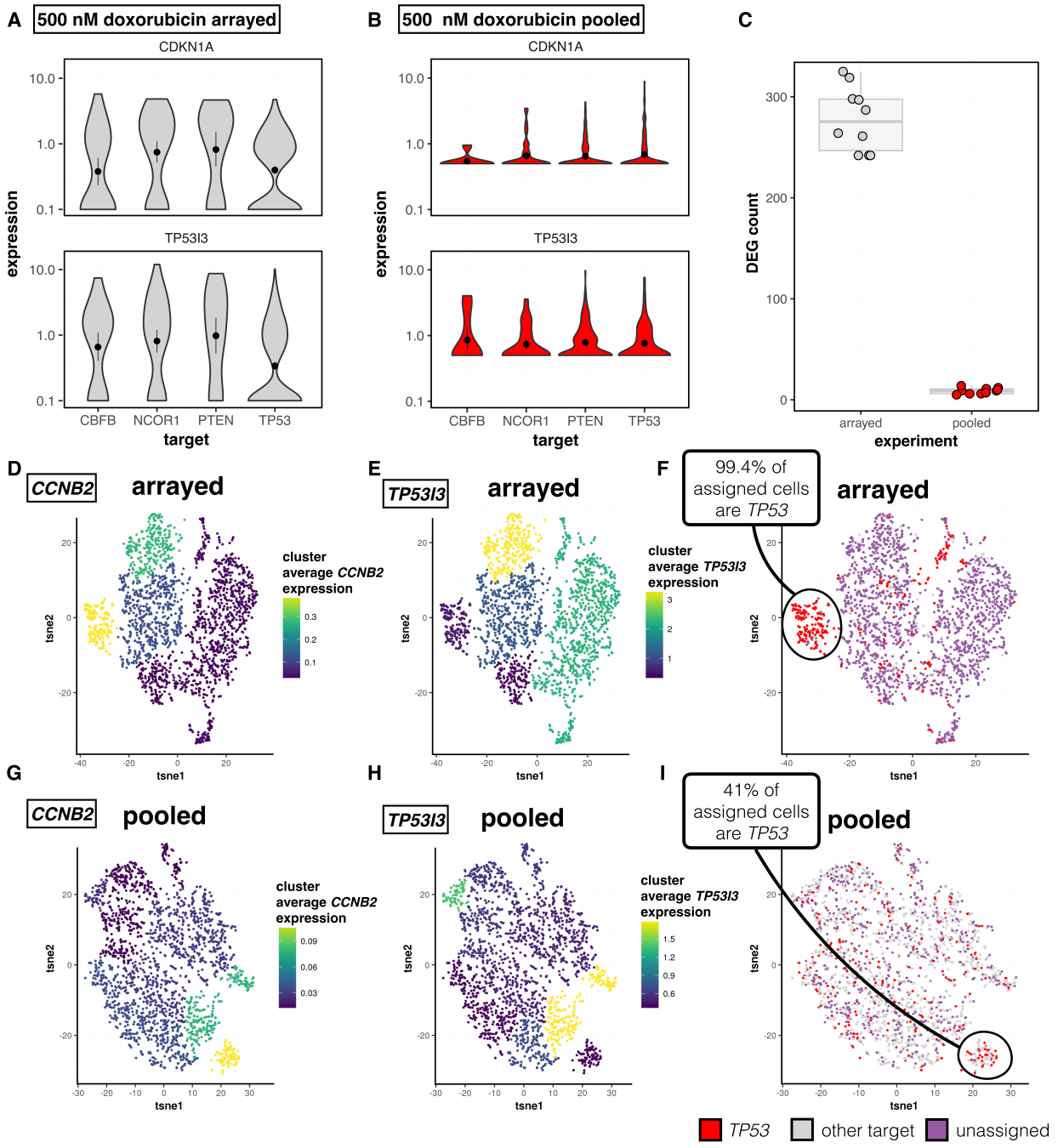


Figure 2.4: Experiments were performed at different times but under the same conditions. The arrayed experiment was performed as a pilot experiment with four targets and observed an overall low rate of cells with detected barcodes. The pooled experiment was performed afterwards with 10 targets and a set of non-targeting controls and we observed a high proportion of cells with detected barcodes and good coverage of the library. (*legend continued on next page*)

(*continued*) To compare these experiments, only the four overlapping targets were considered for differential expression testing and the number of cells containing an sgRNA to each target and sequencing depth were matched between samples to control for power differences. A) Size-factor normalized CDKN1A and TP53I3 expression across TP53 and the three other targets in arrayed screen. B) CDKN1A and TP53I3 expression across TP53 and three other targets in pooled screen that overlap with the arrayed screen. Values for all violin plots capped to a minimum of 0.1 to facilitate plotting. C) Comparison of the number of differentially expressed genes detected at an FDR of 5% for arrayed across the target label in the arrayed and pooled experiments (n = 259 cells per dataset sampled 10 times such that cell counts per target match between datasets; see methods). For each boxplot from right to left the min, 1st quartile, median, mean, 3rd quartile, and maximum are 242.0, 246.8, 275.5, 277.7, 297.8, 325.0; and 5.00, 6.25, 9.00, 8.90, 10.75, 14.00 respectively. D-I show t-SNE embeddings of all cells collected in arrayed and pooled experiments (n = 2274 cells and n = 2191 respectively). D and G) Average size-factor normalized expression of CCNB2 for the arrayed and pooled screens. E and H) Same as D and E but for TP53I3. There is a single distinct CCNB2-high, TP53I3-low cluster one expects for TP53-null cells in this experiment. F and I) The same t-SNE embeddings, colored by their knockout assignments; TP53, other targets, and cells without any assignment are shown as separate colors. The arrayed experiment shows that 99.4% of the cells with an assigned target within the CCNB2-high, TP53I3-low cluster are TP53 knockouts compared to only 41% in the corresponding cluster from the pooled experiment.

from scRNA-seq data alone, this rate was 94% with targeted amplification (Figure 2.9A). In contrast with our original pooled experiment, tSNE embedding of doxorubicin-exposed cells from this experiment yielded a cluster almost entirely composed of cells containing TP53-targeting sgRNAs (Figure 2.9B). Specifically, the 262 cells in this cluster include 90.5% with TP53-targeting guides, 7.6% with guides targeting other genes, 0% with NTC guides, and 1.9% unassigned cells. In contrast, the remaining 5,617 cells include 3.2% with TP53-targeting guides (presumably cells

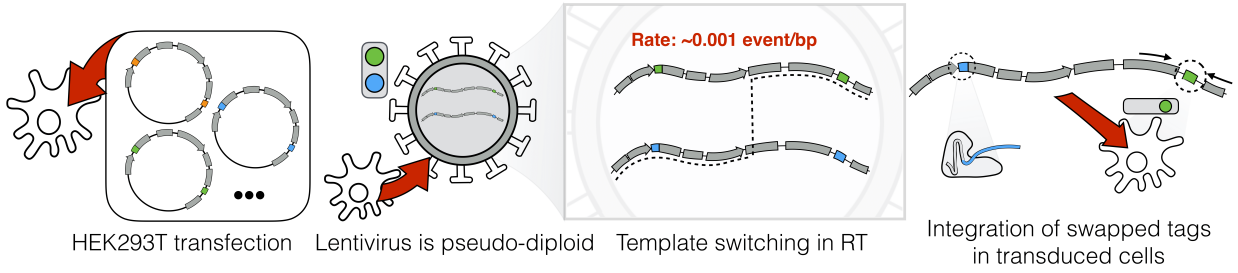


Figure 2.5: Schematic illustrating how template switching in lentivirus leads to the swapping of associations between the sgRNAs and their corresponding barcode.

in which LoF editing failed to occur), 84.2% with guides targeting other genes, 7.5% with NTC guides, and 5.2% unassigned cells. Expression levels of the p53 targets CDKN1A and TP53I3 (el Deiry et al., 1993; Contente et al., 2002) were markedly lower in the TP53-targeted cluster (Figure 2.9C); and 4,277 and 2,186 DEGs (false discovery rate (FDR) 5%) were identified relative to cells with NTC guides in the doxorubicin-treated and untreated (mock) conditions, respectively. Thus, our improved CROP-seq protocol achieves the power and negligible sgRNA swap rate of the arrayed format without sacrificing the scalability of a pooled cloning and lentiviral production workflow.

Upon tSNE analysis of both mock and doxorubicin-treated cells (Figure 2.10A,B), we find several tumor suppressors whose distribution across clusters is significantly different compared to that of NTCs (FDR 5%; 13 and 14 targets with significant changes in the mock and doxorubicin conditions, respectively) (Figure 2.10C-F). We tested for target enrichment within clusters and generated average expression profiles for each enriched target–cluster pair. Gene set enrichment analysis of the most highly loaded genes in the principal components of these average expression profiles show many targets to be associated with increased proliferation and a decreased DNA damage response, most prominently with targeting of TP53 (Figure 2.11).

To further assess the impact of template switching on sensitivity, we permuted target labels within our own CROP-seq tumor-suppressor screen, observing a 2.9-fold reduction in the number

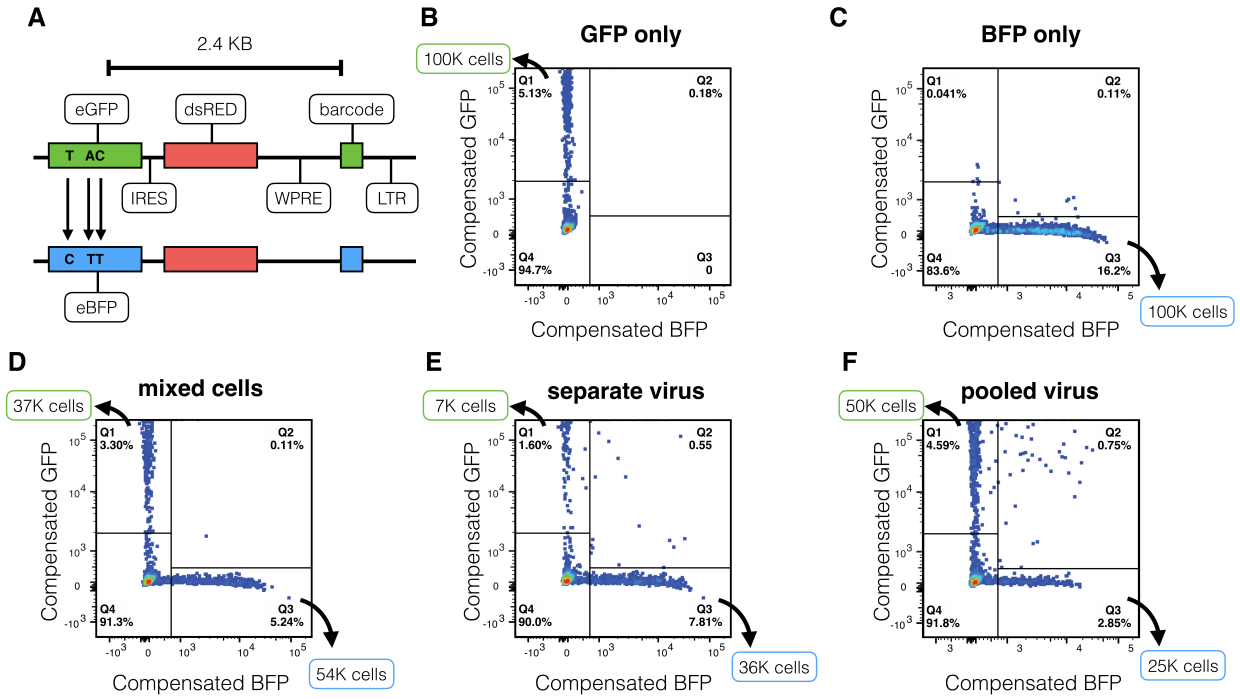


Figure 2.6: Design and sorting of GFP and BFP positive fractions in lentivirus barcode swapping experiment. A) Schematic of vectors (pHAGE-GFP and pHAGE-BFP) designed to quantify template switching rate at 2.4 kb using a FACS readout. FACS plots are shown for sorted cells in samples corresponding to B) GFP only transduced cells C) BFP only transduced cells D) GFP and BFP only transduced cells mixed just prior to FACS as a control E) cells transduced with BFP and GFP virus that was generated separately but pooled prior to transduction F) cells transduced with BFP and GFP virus that was generated from pooled plasmids. The fraction of green plasmids assumed in the determination of lentivirus swap rate from FACS experiments is taken as the fraction of GFP+ cells relative to the total GFP+ and BFP+ cells from this sort (4.59 / (4.59 + 2.85) or 61.7%). This accounts for the fact that plasmids were likely not completely equimolar. The approximate number of total cells cells sorted in each fraction is indicated along the appropriate axes on each plot. Similar results were obtained for B-F in n = 2 independent viral transductions.

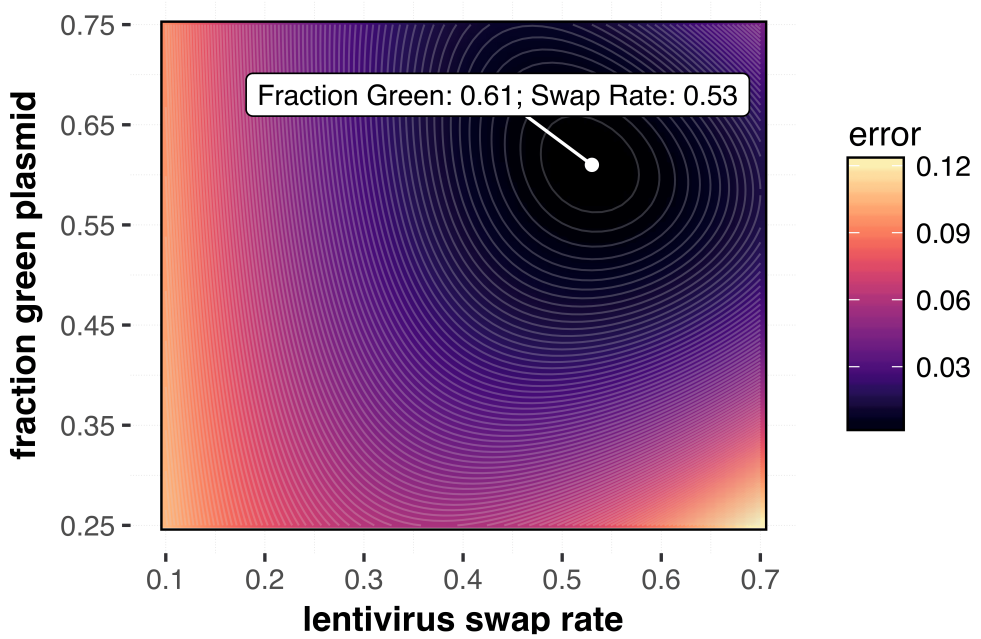


Figure 2.7: Simulation of concordance between observed and expected data obtained from FACS experiment in Figure 2.6 to quantify template switching rate at 2.4 kb separation between paired sequences. Figure 1e assumed a fraction of 0.617 of GFP plasmid in the original green plasmid / blue plasmid mix as determined from FACS in Figure 2.6. In this figure, both the fraction of GFP plasmid and lentivirus swap rate are varied to obtain the set of parameters that best fit the collected fraction GFP and BFP barcodes in the green and blue sorted samples (n = 4 measurements). The sum of the squared error between expected and observed values from FACS given each combination of parameters is shown.

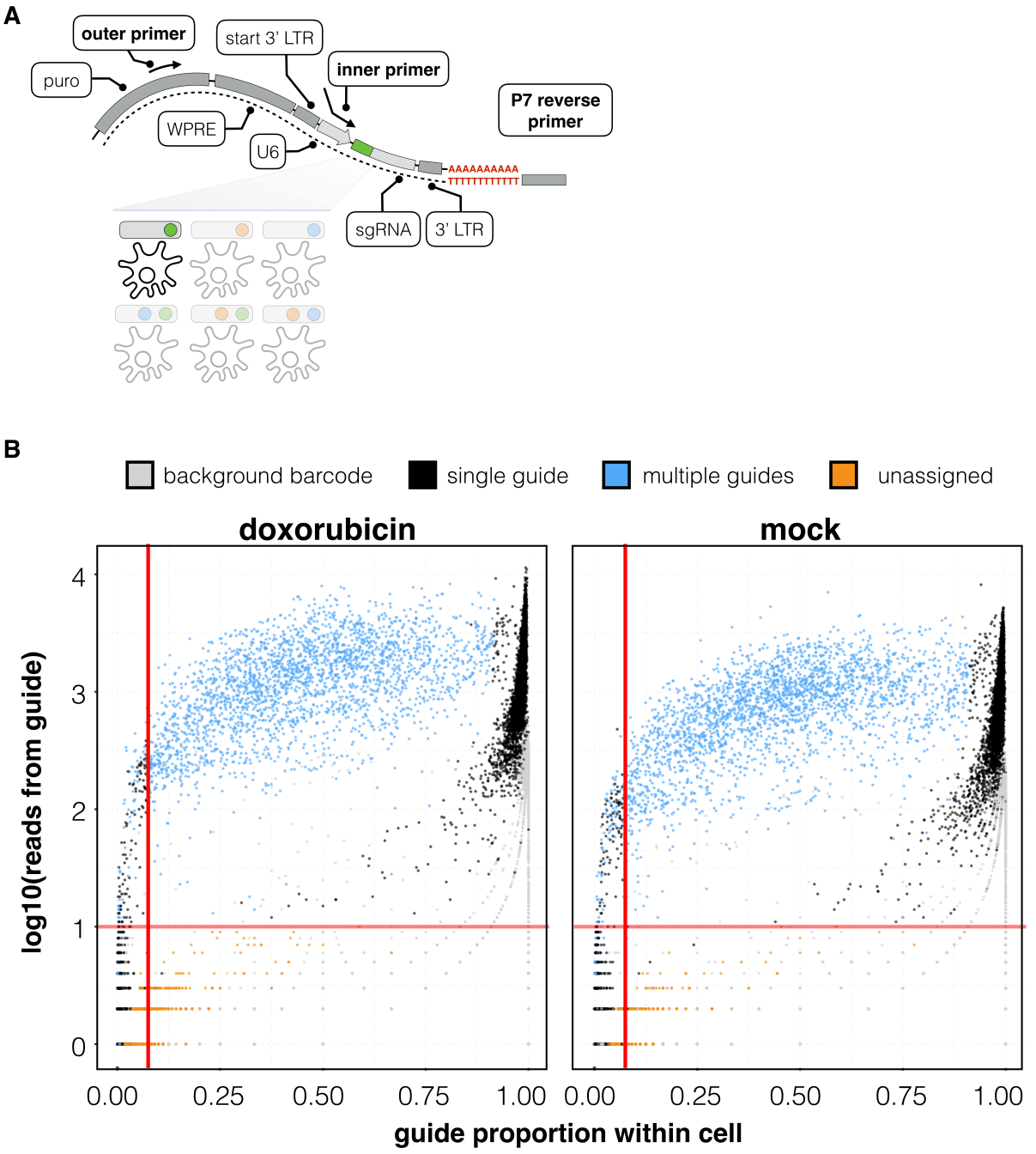


Figure 2.8: CROP-seq PCR enrichment. A) Schematic of PCR enrichment of barcoded transcripts from CROP-seq samples. B) Guide transcript enrichment quality control plot for tumor suppressor knock-out screen performed with CROP-seq. (*legend continued on next page*)

(*continued*) Each dot represents a barcode sequence observed in a given cell. Plot of reads for a given barcode against the proportion of all barcode reads observed in a given cell for every barcode/cell pair colored according to whether the cell barcode is determined to be background from whole transcriptome scRNA-seq data, a cell with a single guide assignment, a cell with multiple guide assignments, or a cell that ultimately receives no assignment given the thresholds applied. Red lines indicate the lower-bound thresholds used to distinguish noise from true guide observations (10 reads and 0.075 proportion within cell). All guide observed above the red lines are assigned to their respective cells. Left, Doxorubicin treated sample from CROP-seq experiment. Right, Mock sample from CROP-seq experiment. Note that the rate of cells with single guide assignments across both experiments above is approximately 81%, which is in line with the 80% expected value at an MOI of 0.45 in a selected population

of DEGs across targets at a swap rate of 50%. The number of significant targets was also reduced, to just 4/13 (TP53, STK11, CHEK1, and NCOR1) and 3/14 (TP53, RB1, and ARID1B) in the mock and doxorubicin conditions, respectively. Additionally, simulations of 50% swapping on the larger (50,000 cells) unfolded-protein response screen from Adamson et al. (2016) with arrayed lentiviral production resulted in a 1.9- and 2.8-fold reduction in the number of DEGs when using 25,000 and 6,000 cells, respectively (Figure 2.12). Altogether, these simulations demonstrate that the reduction in power consequent to swapping is dependent on the number of cells captured, the number of targets, and the effect size of those targets.

Although CROP-seq is not subject to sgRNA-barcode swapping, it is limited by its placement of the sgRNA in the lentiviral LTR, as larger intervening sequences such as dual sgRNA designs16 might render the LTR nonfunctional (Datlinger et al., 2017). To enable incorporation of longer cassettes, we placed the sgRNA cassette between the WPRE and LTR. In this design (pHAGE-scKO), copying of the sgRNA between LTRs would not occur, but the guide sequence would still contribute to overlapping Pol II and Pol III transcripts (Figure 2.13).

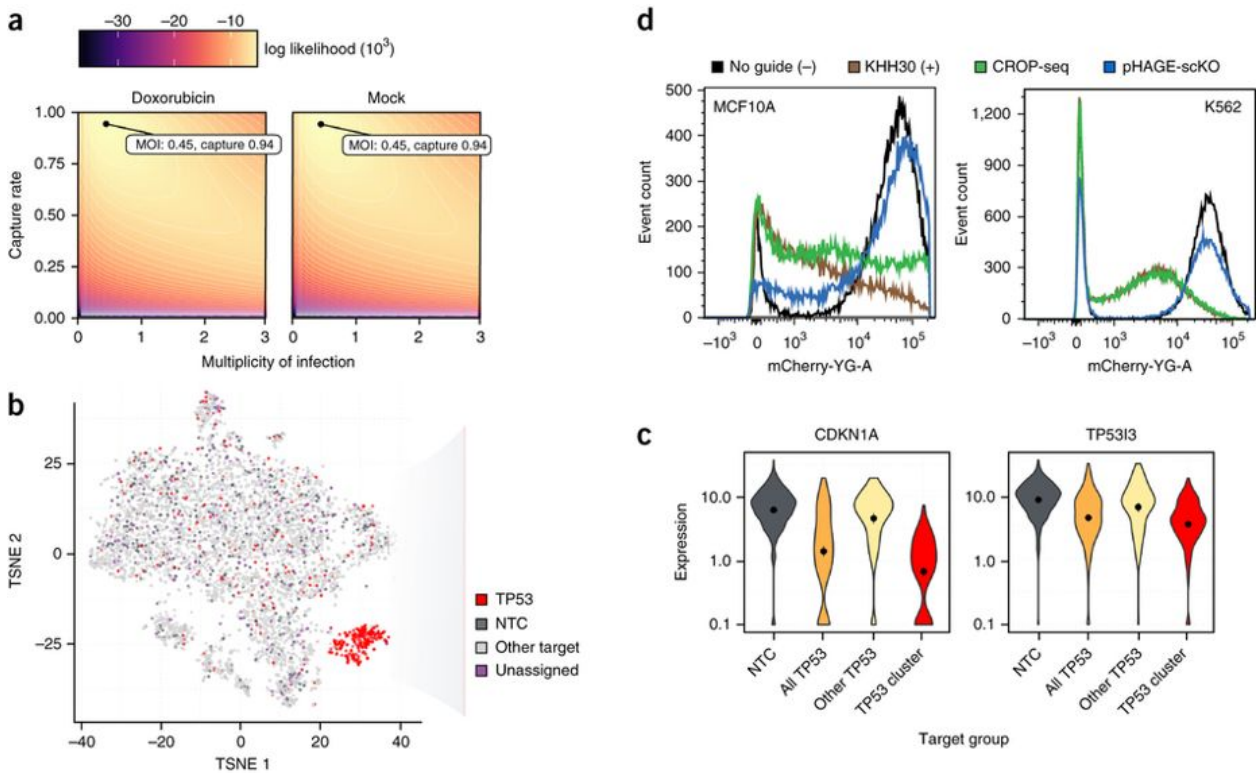


Figure 2.9: CROP-seq with PCR enrichment offers improvements over alternate screen designs in a tumor suppressor knockout screen. A) Most likely multiplicity of infection (MOI) and capture rate of barcoded transcripts in CROP-seq screen based on a generative model. B) tSNE embedding of a doxorubicin-treated sample highlighting cells that contain TP53 guides, nontargeting controls (NTC) or non-TP53 guides, and unassigned cells (n = 5,879 cells). C) CDKN1A and TP53I3 expression in cells expressing either nontargeting controls or TP53 guides. Cells with TP53 guides are further stratified by inclusion in the TP53-enriched cluster from b. Values below 0.1 are not shown to facilitate plotting. D) CRISPRi knockdown of mCherry in MCF10A and K562 cells not expressing a guide (– control), KHH30 (+ control), CROP-seq, and pHAGE-scKO design. All vectors have been modified to contain a CRISPRi-optimized backbone. pHAGE-scKO places the sgRNA within a Pol II 3' UTR and does not knock down mCherry expression.

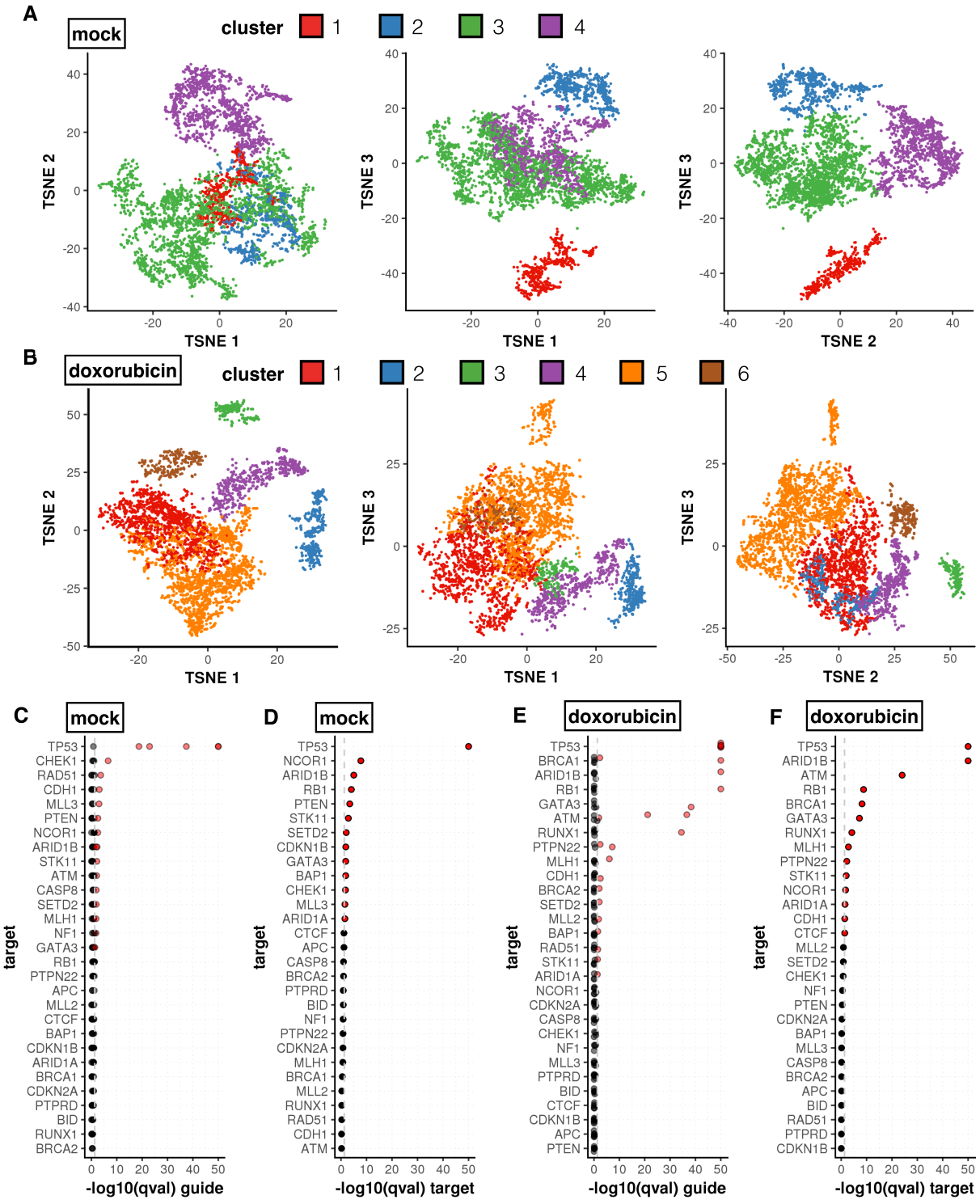


Figure 2.10: Loss of several targets alter the distribution of mock and doxorubicin exposed cells within tSNE clusters. (*legend continued on next page*)

(*continued*) A-B) 3D tSNE embedding and clustering of all cells with single target assignment in mock and doxorubicin treated samples (n = 4740 and n = 4467 cells respectively). C-F) Chi-squared test qvalues resulting from testing for differences in the distribution of targets in our screen at both the individual sgRNA (C and E) and overall target levels (D and F). Comparisons are relative to the distribution of non-targeting controls across tSNE clusters for mock and doxorubicin treated samples, respectively (qvalues were capped to 1e-50 for visualization). Significant differences below a qvalue of 0.05 are colored in red (boundary marked by the grey dashed line).

To evaluate this design, we compared the ability of pHAGE-scKO, CROP-seq, and a standard lentiviral sgRNA expression vector, pKHH030 (Han et al., 2017), all containing a CRISPRi-optimized backbone, to inhibit transcription via CRISPRi, targeting the promoter of an mCherry transgene. Whereas pKHH030 and CROP-seq exhibited efficient inhibition, pHAGE-scKO had poor efficacy (Figure 2.9D). Consistent with this, we observed low editing rates with pHAGE-scKO (88% with pLGB control vs. 29% with pHAGE-scKO). Recent studies suggest interference when Pol II and Pol III transcripts overlap (Lukoszek et al., 2013; Yeganeh et al., 2017). We hypothesize that the poor efficacy of pHAGE-scKO is due to the blasticidin resistance gene inhibiting sgRNA expression. In contrast, CROP-seq likely maintains efficacy because the second integrated copy of the sgRNA (copied to the 5' LTR) does not overlap a Pol II transcript.

CRISPR-based pooled genetic screens coupled to scRNA-seq phenotyping have the potential to be extremely powerful. However, several published designs, and our own initial design, are susceptible to high rates of sgRNA-barcode swapping (diagrams of all relevant vectors are shown in Figure 2.14). Importantly, we do not expect that positive conclusions drawn by published studies using such designs in conjunction with pooled lentivirus production (Xie et al., 2017; Dixit et al., 2016; Jaitin et al., 2016) are incorrect. Each of these studies examined few targets and collected large data sets, raising their baseline sensitivity. However, given the high cost of scRNA-seq and impetus to expand the number of targets in such screens, our observations are highly relevant for

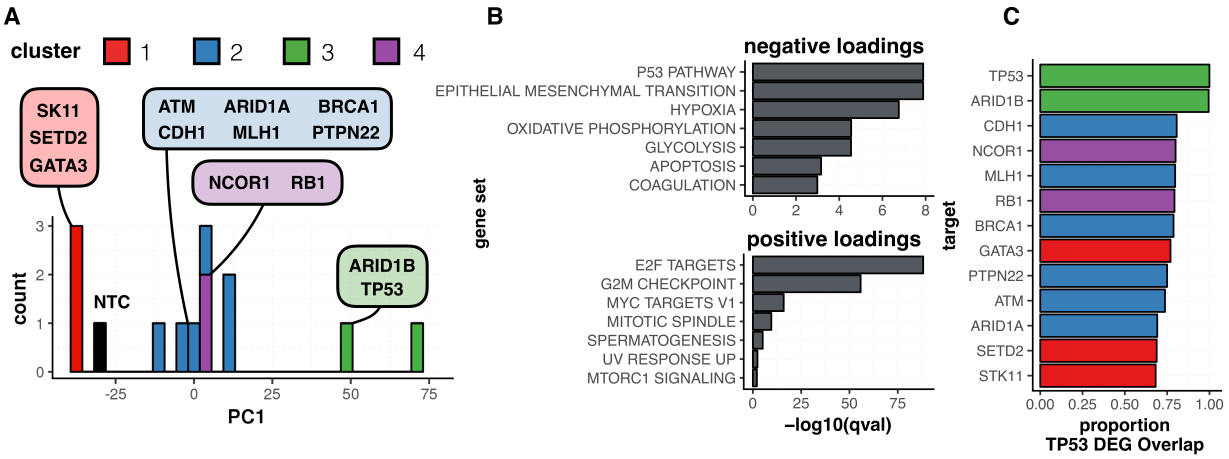
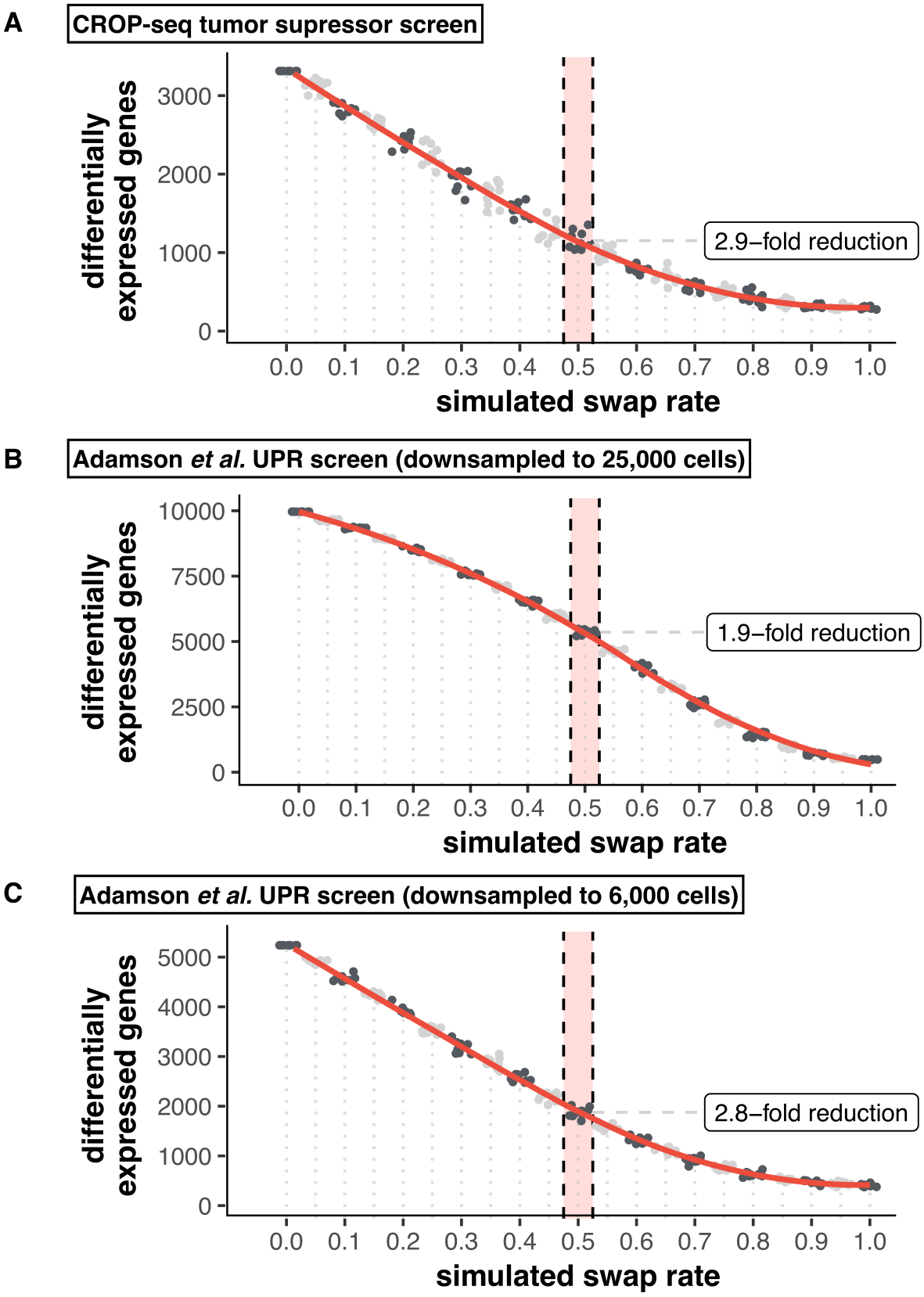


Figure 2.11: Enriched target-cluster pairs highlight tumor suppressors that share various degrees of a TP53 deficient signature. A) Fisher's exact with weights applied to guides according to an expectation maximization procedure were performed for the doxorubicin treated sample to find clusters from Figure 2.10 panel B where particular targets were found to be enriched. Cells with target-cluster pairs that showed enrichment were used to generate an aggregate expression profile for every target within genes that are differentially expressed between TP53 and non targeting controls (NTC). A PCA was performed on these average expression profiles and a distribution of targets across PC1 is shown colored by the cluster in which they were found to be enriched. B) Gene set enrichments for top positively and negatively loaded (less than -0.02 or greater than 0.02; n = 491 and n = 347 genes respectively) genes along PC1 (qval $<$ 0.01). C) Differential expression tests were performed for cells within each enriched target-cluster pair, comparing each target to all NTC cells. The proportion of overlap between these differentially expressed genes and the genes differentially expressed between TP53 and NTC is shown.



(*legend on next page*)

Figure 2.12: Swap rate simulations for our own CROP-seq tumor suppressor screen and the unfolded protein response screen from Adamson et al. (2016) Each dataset was subjected to simulation of progressively higher fractions of target assignment swapping to mimic the impact of template switching. Number of differentially expressed genes across the target label at FDR of 5% is plotted at each swap rate for 10 samplings per rate. 0.5 corresponds to the 50% swap rate determined via FACS. A) CROP-seq tumor suppressor screen from our study (doxorubicin treated condition; n = 4467 cells used in tests). B) Unfolded protein response screen from Adamson et al. (2016) downsampled from 50,000 to n = 25,000 cells to make simulations computationally feasible. C) Unfolded protein response screen from Adamson et al. downsampled to n = 6000 cells to illustrate how reduced power impacts the observed impact from simulated swapping.

future studies. Reductions in power may be partly overcome by filtering cells that appear inconsistent with their assigned target (Dixit et al., 2016), or completely overcome with arrayed lentiviral production (as in Adamson et al. (2016)). However, computational filtering has the potential to introduce biases, and itself reduces power by discarding collected data, while arrayed lentiviral production dramatically limits scalability.

A viable alternative is the recently published CROP-seq method (Datlinger et al., 2017). By coupling targeted sgRNA amplification and CROP-seq, we doubled the proportion of cells in which guides are assigned to 94%. The attractive features of this approach include the simplicity of the cloning protocol, its compatibility with lentiviral delivery, the high rate of recovery of sgRNA-cell associations, and minimized risk of template switching.

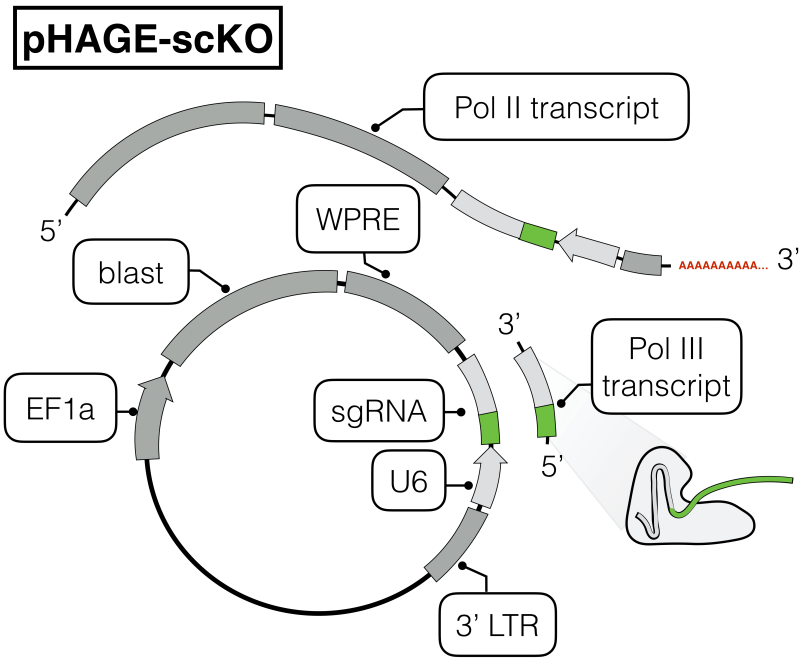
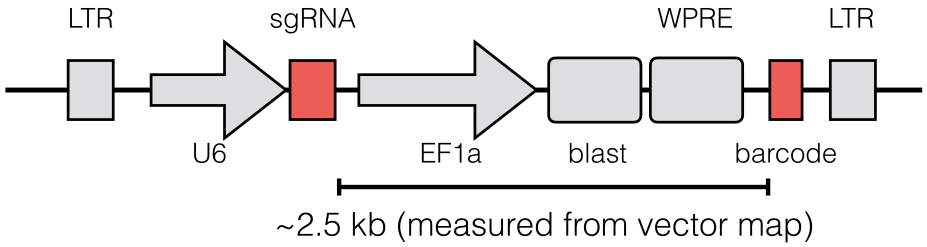
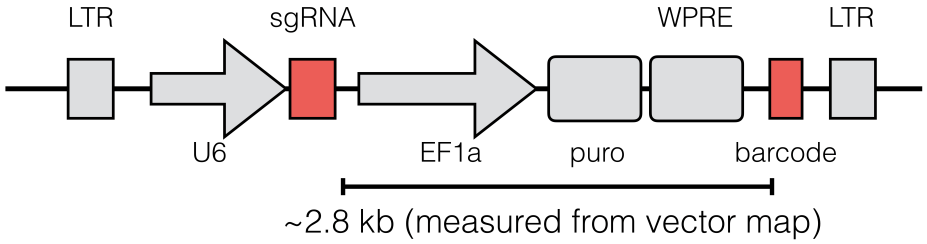


Figure 2.13: Schematic of pHAGE-scKO design where guide is placed in the 3' UTR outside of the LTR such that a second copy is not generated during integration.

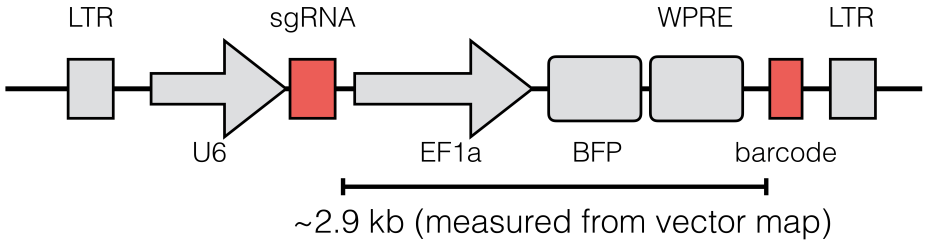
A **pLGB-scKO (our initial design)**



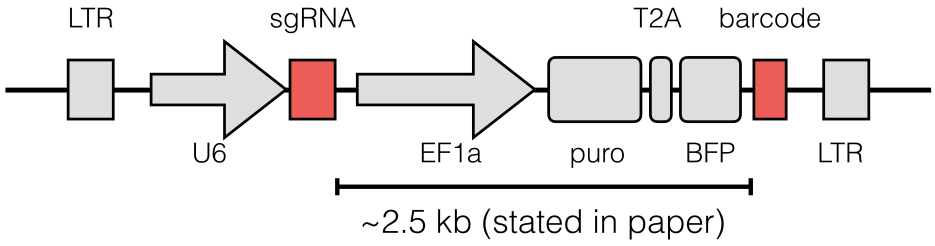
MOSAIC-seq



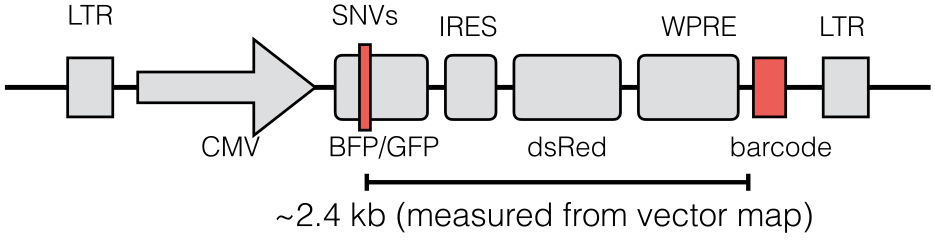
CRISP-seq



Peturb-seq



B **pHAGE-GFP/BFP (our vector to measure swap rate)**



(*legend on next page*)

Figure 2.14: A comparison of all relevant vector designs that rely on the linkage of sgRNAs and distal barcode. A) vector maps for our own initial design (pLGB-scKO) and all three published designs that rely on linkage between an sgRNA and a distal barcode. Distances between the guide sequence and the barcode are indicated in the schematic of each vector. B) Vector map for pHAGE-GFP/BFP, the vector we used to quantify the rate of swapping that results during lentiviral packaging at a distance of 2.4 kb or greater.

2.3 METHODS

2.3.1 Cell culture

MCF10A immortalized breast epithelial cells (Debnath et al., 2003) were purchased from ATCC and cultured in DMEM/F12 (Invitrogen) supplemented with 10% FBS, 1% pen–strep, 10 ng/mL EGF, 1 μg/mL hydrocortisone, 5 μg/mL insulin, and 100 ng/mL cholera toxin. K562 cells were cultured in RPMI 1640+L-Glutamine (Gibco) supplemented with 10% FBS (Rocky Mountain Biologicals) and 1% pen–strep (Gibco).

2.3.2 Generating inducible Cas9-expressing MCF10A cell lines

Lentivirus containing either a doxycycline-inducible or constitutively expressed Cas9 construct were produced by transfecting 293T cells with either pCW-Cas9 (Addgene 50661) or lentiCas9-Blast (Addgene 52962) using the ViraPower Lentiviral Expression System (Thermo) according to manufacturer's instructions. 48 h post-transfection, supernatant was collected and debris removed using a 40 μm syringe filter. MCF10A were transduced with viral supernatant for 48 h and selected with 1 μg/mL puromycin (pCW-Cas9) or 10 μg/mL blasticidin (lentiCas9-Blast) for 96 h. For cells expressing a doxycycline-inducible Cas9, single-cell clones of MCF10A-Cas9 cells were generated by dilution, clones were expanded, and Cas9 expression was confirmed by immunoblotting 96 h following addition of doxycycline at 1 μg/mL. lentiCas9-Blast cells were maintained as

a polyclonal line.

pCW-Cas9 cells were used for initial arrayed and pooled screens as well as quantification of editing rates in pHAGE-scKO vector. lentiCas9-Blast cells were used for all CROP-seq experiments.

2.3.3 Initial tagged transcript cloning method

Because of high rates of barcode–sgRNA swapping when using this design, we do not recommend use of this protocol.

LentiGuide-puro (Addgene 52963) was modified to confer blasticidin resistance. Puro and its EF-1A promoter were removed via double digest with NEB SmaI (8 h at 25 C) and MLU1-HF (8 h 25 C). This product was gel purified using QiaQuick Gel Extraction kit (Qiagen). EF-1A promoter and blasticidin, each with 20 bp homology on both ends, were prepared via PCR from lentiCas9-Blast and gel purified. Fragments were assembled into digested lentiGuide-puro vector using the NEBuilder HiFi DNA Assembly kit with inserts in two-fold molar excess and transformed into NEB C3040H E. coli and allowed to incubate overnight at 30 C. Clones were picked from plate, allowed to grow in LB + amp overnight at 30 C, and purified using Qiagen Miniprep kit. Individual clones were validated via Sanger sequencing.

Lentiguide-blast was linearized using a digest with BsmB1 (Thermo) at 37 C for 5 h followed by digestion with SalI HF (NEB) overnight and gel purification. Oligos containing guide sequences and their corresponding barcodes were designed according to the following:

tGTGGAAAGGACGAAACACC[G][guide] gttttagagctaGAAAtagcagagacgCGTCTCAgatctc-cctttgggccgcctccccgcg [barcode]tcgactttaagaccaatgacttaca

Where [guide] is a 20 bp guide sequence and [barcode] is an 8 bp barcode sequence uniquely paired to an sgRNA. The [G] included before guide is required for expression from Pol III promoters. Guides/barcodes that generate an extra BsmB1 restriction site when used in this design were excluded. RUNX1 only included four guides because of this filter.

A library of these oligos was ordered as Ultramers from IDT. All oligos were resuspended in

water, pooled at equimolar concentrations, and amplified using a 50 μl KAPA HiFi HotStart Ready Mix PCR reaction with 1ng of input DNA. The resulting product was cleaned with a Zymo DNA Clean and Concentrator kit. The purified inserts were assembled into linearized lenti-Guide-blast using the NEBuilder HiFi DNA Assembly kit and a molar excess of 1:5 vector to insert. Assembled products were transformed into NEB C3040H E. coli and grown overnight at 30 C in LB + amp. Product was prepared using a plasmid Miniprep kit (Qiagen).

To prepare the insert for the final reaction, a region from the backbone sequence for the CRISPR sgRNA to a region toward the end of the WPRE element was amplified using the KAPA HiFi Hotstart Master Mix and purified using the Zymo Clean and Concentrator kit. The primers used in this reaction add BsmB1 cut sites that generate complementary ends in the final cloning step following digestion. This amplified fragment was ligated into PGEM-T using the PGEM-T kit a clone selected and validation of individual clones by Sanger sequencing. The validated construct was digested with BsmB1 (Thermo) and gel purified.

The fragment isolated from PGEM-T was then ligated into the linearized vector using a 3:1 molar excess of insert to vector using T4 DNA Ligase (New England Biolabs) and overnight incubation at 16 C. Ligation products were transformed into NEB C3040H (stable) competent cells and grown overnight at 30 C in LB + amp. Plasmids were recovered using a Plasmid Miniprep kit (Qiagen).

2.3.4 pHAGE and CROP-seq vector cloning

The pHAGE_dsRed_IRES_zsGreen vector was modified to contain a multiple cloning site as described in ”Quantification of template switching in lentivirus packaging using FACS.” The U6-sgRNA cassette containing a 500 bp filler removable by Bsmb1 digest was ordered as an IDT gblock. Using the multiple-cloning site, the U6-sgRNA cassette was added in the 3’ UTR of the zsGreen/dsRed transgene via Gibson assembly. This vector was modified to remove the zsGreen/IRES/dsRed cassette and replace the CMV promoter with an EF1α promoter.

To clone libraries for this vector or CROP-seq vector (Addgene 86708), the starting vector

was digested following the protocol outlined in Sanjana et al. (2014). Oligos corresponding to individual guides with homology for Gibson assembly were ordered as standard DNA oligos from IDT with the following design:

[GCCTTATTTTAACTTGCTATTTCTAGCTCTAAAAC] [GUIDERC][C] [GGTGTTTCGTCCTTTCCACAAGAT]

GUIDERC refers to the reverse complement of the guide sequence. The entire construct may also be reverse complemented, allowing the guide sequence itself to be used rather than the reverse complement.

All oligos were resuspended in water, pooled at equimolar concentrations, and amplified using a 50 μl KAPA HiFi HotStart Ready Mix PCR reaction with 1 ng of input DNA. The following primers were used for amplification:

Forward: 5-GCCTTATTTTAACTTGCTATTTCTAGCT-3 Reverse: 5-ATCTTGTGGAAAGGACGAAACA-3

These reactions were cleaned with a Zymo DNA Clean and Concentrator kit and cloned into the Bsmb1-digested pHAGE vector backbone using the Clontech Infusion HD Cloning Kit. Ligations were performed using 10 fmol of vector and 200 fmol of double-stranded oligo (1:20 molar ratio of vector to insert). Ligation products were transformed into NEB C3040H (stable) cells according to manufacturer recommendations. Transformations were diluted with 250 μL of LB and spread onto 6 LB-AMP plates and incubated at 30 C for 24 h. Colonies were then scraped into LB, bacterial pellet was collected, and plasmids were recovered using a Plasmid Midiprep kit (Qiagen).

The CROP-seq vector with optimized backbone (CROP-seq-opti was cloned in a manner similar to the standard CROP-seq vector but with different homology.

Oligos were ordered with the following 3' homology: 5-gtttAagagctaTGCTGGAAACAGCAtagcaagt-3

If ordering in the same format as above (where the oligo is the reverse complement), GCCTTATTTTAACTTGCTATTTCTAGCTCTAAAAC, would be replaced by the reverse complement of the above sequence and amplified with primers: Forward: 5-atcttGTGGAAAGGACGAAACA-3

Reverse: 5-acttgctaTGCTGTTTCCAGC-3

Each of these vectors is also compatible with alternative cloning protocols for lentiGuide-Puro vectors (as long as any homology is adjusted as needed).

2.3.5 Quantification of template switching in lentivirus packaging using FACS

A multiple-cloning site was cloned into pHAGE_dsRed_IRES_zsGreen lentiviral vector between the WPRE and 3 LTR. The multiple-cloning site was assembled from annealing and extension of WPRE_MCS_insert_W and WPRE_MCS_insert_R: WPRE_MCS_insert_W: 5-ctttgggccgcctccccgcctgggcgcgccATA
3

WPRE_MCS_insert_R: 5-cagctgccttgtaagtcattggtcttaaaggctcgagCCATCAgctagcTGTTATgg-3

The plasmid was amplified by inverse PCR with pHAGE_WPRE_MCS_GIBS_F and R:

pHAGE_WPRE_MCS_GIBS_F 5-TGGctcgagcctttaagaccaatgacttacaaggcagctg-3

pHAGE_WPRE_MCS_GIBS_R 5-ctagcTGTTATggcgcgcccaggcggggaggcggcccaaag-3

The fragments were cloned by Gibson assembly. Clones of pHAGE_dsRed_IRES_zsGreen_WPRE_MCS were chosen by Sanger sequencing and expression of the fluorescent proteins after transfection and lentiviral packaging.

To make pHAGE EBFP or EGFP_IRES_dsRed_WPRE_MCS, pHAGE_dsRed_IRES_zsGreen_WPRE_MCS was cut with BamHI and ClaI to remove the zsGreen and IRES. The ends were blunted and re-ligated to make pHAGE_dsRed_WPRE_MCS. EGFP or EBFP (amplified with eGFP_gibsF and eGFP_IRES_GibsR) and an IRES (IRES_GibsF, IRES_GibsR) were cloned into the NotI site 5' of the dsRed by Gibson assembly. EBFP was ordered as a gblock from IDT with 3 nt changes from EGFP. Correct clones were identified by sequencing. The dsRed is not expressed in this construct.

eGFP_gibsF: 5-gccatccacgctgttttgacctccatagaagacaccggcATGGTGAGCAAGGGCGAGGAG-
3

eGFP_IRES_GibsR: 5-ggatccCTACTTGTACAGCTCGTCCATGCCG-3

IRES_GibsF: 5-ATCACTCTCGGCATGGACGAGCTGTACAAGTAGggatccctccccccccccctaacgttac-
3

IRES_GibsR: 5-ctccttgatgacgtcctcggaggaggccatggcggccatgtgtggccatattatcatcgtgtttttcaaagg-3

EBFP 5-ATGGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGA
3

15 bp barcodes (lenti-barcode and lenti-barcode-r) were cloned into the multiple-cloning site between the WPRE and 3' LTR for both the EBFP and EGFP constructs by Gibson assembly. Single clones were prepared and the barcode identified by Sanger sequencing.

lenti-barcode: 5-atctccctttgggccgcctccccgcctgggGGATCCAGNNNNNNNNNNNNNNNtcgagcctttaagaccaatg
3

lenti-barcode-r: 5- CCTTGTAAGTCATTGGTCTTAAAGGCTCGA -3

Lentivirus was packaged by transfection of barcoded EGFP or EBFP constructs either alone or in an equimolar mix along with helper plasmids (pHDM-Hgpm2, pHDM-Tatlb, pRC-CMVRev1b, and pHDM-VSV-G) into HEK293T cells using Lipofectamine 2000 (Invitrogen). Viral supernatant was collected after 48 h, spun to remove debris, snap frozen in liquid nitrogen, and stored at -80 C. To titer the packaged lentiviruses, they were thawed on ice and added to MCF10A cells with media containing 8 μg/ml polybrene, and the frequency of transduced cells 48 h post-transduction was determined by flow cytometry.

To sort blue+ and green+ populations, 400,000 of MCF10A TP53 cells (Horizon Discovery) in 5 ml media plus 8 μg/ml polybrene were transduced at a MOI ˜0.1, with either of the EGFP or EBFP expressing viruses that had been packaged singly, a mix of the EGFP and EBFP expressing viruses that had been packaged singly, or the EGFP and EBFP expressing viruses that had been packaged together. The cells were cultured for 4 weeks to avoid residual plasmid contamination following transduction. An equal number of cells transduced with EGFP and EBFP virus were mixed to determine the rate of contamination resulting from FACS error. The mixed cells along with others were sorted for blue+ or green+ populations using a FACS Aria II (Becton Dickinson) that had been compensated for the overlap between the EBFP and EGFP emission spectra. Genomic DNA was harvested from each population using the Qiagen DNeasy kit, and barcodes were amplified from 2-36 ng of genomic DNA in 50 μl Robust polymerase (Kapa) reactions with primers bwds_p5_WPRE_BC_F and bwds_next_WPRE_BC_R.

bwds_next_WPRE_BC_R: GGCTCGGAGATGTGTATAAGAGACAG

5-gaaatcatcgtcctttccttggct-3

bwds_p5_WPRE_BC_F: 5-AATGATACGGCGACCACCGAGAgcgccgatgccttgtaagtcattggtcttaaaggctc-3

PCR products were purified with Ampure (Agilent) and P7 index sequences added by an additional six cycles of PCR. PCR products were purified, quantified, pooled, and single-end sequenced on an Illumina Nextseq500 with Read1 primer bwds_WPRE_bc_seqF and standard Illumina i7 primers.

bwds_WPRE_bc_seqF: 5-GCGCCGATGCCTTGTAAGTCATTGGTCTTAAAGGCTCGA-3

2.3.6 Analysis of FACS data from pHAGE-GFP and pHAGE-BFP experiments

Background percentage of contaminating barcodes in the BFP/GFP-sorted cells from the mixed cells control was subtracted from numbers obtained for the pooled virus samples. Fraction of GFP cells, determined from FACS gating, was fixed; and the expected fraction of barcode contamination in the BFP and GFP was simulated. Note that the expected contamination of green barcodes in the BFP-sorted cells is the template-switching rate multiplied by the fraction of green cells. The expected rate of contamination of BFP barcodes in the GFP-sorted cells is the template-switching rate multiplied by the BFP fraction (1 - GFP fraction). Sum of the squared error between observed and expected values for rates of contamination was calculated for a range of different lentivirus swap rates, and minimal value was taken to be the most likely swap rate.

Note that, unlike a library of plasmids, in a mix of two plasmids, only half of all chimeric products will be detectable as many virions will be homozygous (i.e., contain the same construct, and thus chimeric products are identical to the original). To give an analogous example, in a barnyard experiment for a single-cell assay, mouse–mouse or human–human multiplets cannot be detected and thus estimated rates of 'doublets' have to be adjusted accordingly. When the plasmids are equimolar and the swap rate is 50%, for example, one would expect to observe a 75% rate of the intended barcode and a 25% rate of the unintended barcode. This ratio will change according to the molar concentration of the two plasmids. In Figure 1e, we assume the pool was composed

of 61.7% GFP plasmid, corresponding to the fraction of GFP++ cells relative to the total number of GFP++ and BFP++ cells 4.59/(4.59 + 2.85) or 61.7% as explained in Figure 2.6. This analysis was also performed without fixing the fraction of GFP++ cells to the value measured by FACS to ensure results were concordant (Figure 2.7). The minimum sum of squared error over the grid of simulated lentivirus swap rate and fraction of GFP cells were taken to be the most likely set of parameter values.

2.3.7 CRISPRi experiment

K562 expressing dCas9-BFP-KRAB (gift of the Bassik lab, Addgene 46911) and MCF10A expressing dCas9-BFP-KRAB (made by transduction with lenti_UCOE_EF1-dCas9-BFP-KRAB, plasmid, a gift of the Weissman lab (available on Addgene soon; see `https://weissmanlab.ucsf.edu/CRISPR/CRISPRiacelllineprimer.pdf`) were transduced with lenti-mCherry under control of a CAG promoter (pCAG_mCherry pKH143, gift of the Bassik lab, unpublished), and sorted such that the resulting population is enriched for mCherry expression.

A spacer targeting the CAG promoter was cloned into the KHH030 (Addgene 89358), CROP-seq, and pHAGE-scKO sgRNA expression vectors. The CROP-seq and pHAGE-scKO vectors were modified by Q5-Site Directed Mutagenesis (NEB) to use the previously described sgRNA-(F+E)-combined optimized backbone (Chen et al., 2013) (we refer to this as CROP-seq-opti). The CRISPRi mCherry++ K562 and MCF10A cells were transduced with the CAG-targeting sgRNA and assayed for mCherry.

All viruses for the CRISPRi experiments were made by the Co-operative Center for Excellence in Hematology Vector Production core. All sorting was performed on a FACS Aria II (Becton Dickinson).

2.3.8 Editing-rate experiment for pHAGE-scKO

To confirm that our pHAGE-scKO vector exhibited reduced editing efficiency, we performed editing with a guide to TP53 from our screen (GAGCGCTGCTCAGATAGCGA) in both lentiGuide-

Blast and pHAGE-scKO using our pCW-Cas9 MCF10A cells. Cells were passaged for 18d after induction of Cas9 expression with dox, and gDNA was harvested using Qiagen DNeasy kit and amplified using primers CTAAATGGCTGTGAGAGAGCTCAGCCACACGCAAATTTCCTTCC and ACTTTATCAATCTCGCTCCAAACCCCCTGCCCTCAACAAGATGT. These were then amplified using KAPA HiFi Hotstart Ready Mix (KAPA) using the following indexed primers: AATGATACGGCGACCACCGAGATCTACACacgtaggcCTAAATGGCTGTGAGAGAGCTCAG

CAAGCAGAAGACGGCATACGAGAT[INDEX]gaccgtcggcACTTTATCAATCTCGCTCCAAACC

Libraries were sequenced on MiSeq, and reads were then processed using the method described in McKenna et al. (2016).

Briefly, reads are trimmed of low-quality bases using Trimmomatic, merged using Flash, aligned to the reference of the locus surrounding the guide using needle, and unique genotypes are quantified. The wild-type genotype fraction was taken to be the proportion of non-wild-type alleles. We did not use UMIs in this experiment, and thus it may overestimate editing rate.

2.3.9 Knockout experiments

For all screens, each plasmid library was transfected along with plasmids provided with the ViraPower Lentiviral Expression into 293T cells. At 48 and 72 h post-transfection, supernatant was collected and filtered using a 40 μm steriflip filtration system (EMD Millipore). For arrayed experiments, individual plasmids were transfected and viruses produced as described above. For pHAGE-scKO and arrayed/pooled pLGB-scKO vector experiments, virus was concentrated using Peg-it virus concentration solution (SBI). Viral titer of the concentrated lentiviral library was determined by transduction of MCF10A-Cas9 cells for 48 h at several viral dilutions, splitting cells into replica plates, and subjecting replica plate to blasticidin. Percent control growth was used to assess MOI. MCF10A-Cas9 cells with estimated MOIs of 0.3 carried forward.

For pHAGE-scKO and arrayed/pooled pLGB-scKO vector experiments, media were switched to 1 μg/mL doxycycline to induce expression of Cas9 in pCW-Cas9 cells. LentiCas9-Blast cells were used for CROP-seq experiments. Editing was allowed take place for 14 d for arrayed and

pooled pLGB-scKO and 21 d for pHAGE-scKO and CROP-seq experiments. Media were changed every 48 h, and cells were cultured every 96 h. For the first half of editing, cells were cultured in the presence of 5 μg/mL blasticidin and 0.5 μg/mL puromycin to ensure high sgRNA and Cas9 expression. In all CROP-seq KO experiments (but not our CRISPRi experiment), we used the CROP-seq vector from Datlinger et al. (2017) without modification (Addgene 86708).

2.3.10 Doxorubicin treatment

After editing, MCF10a cells were seeded in 10 cm plates plates at 1 × 106 cells per well, allowed to attach overnight, and media replaced with MCF10A media alone (mock) or MCF10A media containing 500 (arrayed and pooled pLGB-scKO experiments) or 100 nM (pHAGE-scKO and CROP-seq experiments doxorubicin prepared from a 500 μM stock of doxorubicin (Sigma) in water. 24 h after drug exposure, untreated and doxorubicin-treated cells were harvested by trypsinization, washed with PBS, and used for downstream assays.

2.3.11 Single-cell RNA sequencing

Cells were captured using one lane of a 10X Chromium device per sample using 10X V1 Single Cell 3'-Solution reagents (10X Genomics). Approximately 4,000–7,000 cells were captured per lane for each condition. Protocols were performed according to protocol, holding 10–30 ng of full-length cDNA out of downstream shearing and library prep steps in order to provide material for barcode-enrichment PCR.

Final libraries were sequenced on NextSeq500. 10X V1 samples were sequenced using the following read configuration:

R1: 64, R2: 5, I1: 14, I2: 8

Our initial arrayed and pooled doxorubicin-treated samples using pLGB-scKO were aggregated using cellranger aggregate to normalize the average number of mapped reads per cell. This yields an average of 37,732 reads per cell; 2,263 median genes per cell; and a median of 8,279 UMIs per cell.

Our CROP-seq mock sample was sequenced to an average depth of 120,797 raw reads per cell in 6,598 cells. A median of 4,619 genes per cell was detected and a median UMI count of 22,495 per cell. Our CROP-seq doxorubicin-treated sample was sequenced to an average depth of 123,445 raw reads per cell in 6,283 cells. A median of 3,500 genes per cell was detected, and we observed a median UMI count of 15,324 per cell. At this depth the average duplication rate is approximately 78%.

2.3.12 Enrichment PCR

For all experiments, a heminested PCR starting from 5 ng of full-length cDNA was used to enrich for barcodes that assign a target to each cell. PCR reactions were performed with a P7 reverse primer (as introduced by the 10X Chromium V1 oligo DT RT primer). Importantly, the protocol for the 10X V2 protocol (not used here) would be different – see `https://github.com/shendurelab/single-cell-ko-screens#enrichment-pcr` for more information. For pHAGE-scKO and pLGB-scKO, the first PCR was performed with:

5-TCCTGGGATCAAAGCCATAGT-3

and for CROP-seq: 5-TTTCCCATGATTCCTTCATATTTGC-3

as the forward primer, priming to the blasticidin transcript with no nontemplated sequence for pHAGE-scKO and pLGB-scKO, and to part of the U6 promoter in CROP-seq. For pLGB-scKO the second PCR was performed with: 5- TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGACGAGTCGGATC
3

for pHAGE-scKO with: 5-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAACGGACTAGCCTTATT
3

and for CROP-seq with: 5-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGcTTGTGGAAAGGACGAA
3

as the forward primer, priming adjacent to the barcode/guide sequence in each design and adding the standard Nextera R1 primer. Samples were indexed in a final PCR using standard Nextera P5 index primers of the form:

5-AATGATACGGCGACCACCGAGATCTACAC[8bp Index]TCGTCGGCAGCGTC-3

PCRs were cleaned with a 1.0X AmpureXP cleanup and 1 μl of a 1:5 dilution of the first PCR and 1:25 dilution of the second PCR were carried in each reaction.

2.3.13 Digital gene expression quantification

Sequencing data from each sample was processed using cellranger 1.3.1. Each lane of cells was processed independently using cellranger count, aggregating data from multiple sequencing runs. For the comparison between arrayed and pooled screens, cellranger aggregate was used to down-sample data from each screen to an equal average number of mapped reads.

2.3.14 Assigning cell genotypes

Barcode-enrichment libraries were separately indexed and sequenced as spike-ins alongside the whole-transcriptome scRNA-seq libraries. Final UMI and cell-barcode assignments were made for each read by processing these samples with cellranger 1.3.1, as was done for the whole-transcriptome libraries.

A whitelist of guide or target barcode sequences was constructed using all guides or target barcodes in the library. For each read in the position-sorted BAM file output by cellranger 1.3.1, the final cell barcode and UMI are extracted. If either of these fields is not populated, indicating a problem with the sequence, the read is ignored. Using the cDNA read, we attempt to find a perfect match for the sequence preceding the guide or barcode (GTGGAAAGGACGAAACACCG for CROP-seq and CGCCTCCCGCG for pLGB-scKO). If a perfect match is not found, we attempt to locate the sequence using a striped Smith–Watterman alignment. If a match or alignment is found, the guide or barcode sequence is extracted. If the extracted sequence does not perfectly match a whitelist sequence, we search for a matching whitelist sequence within an edit distance of half the minimum edit distance between any pair of guides or barcodes in the library (rounded down). If no match is found, the molecule is ultimately discarded. Matches to the whitelist are tracked for each cell.

We also remove likely chimeric sequences using the approach outlined in Dixit (Dixit, 2016). Briefly, within each cell we calculate the number of times a given UMI is observed with each observed guide assignment. We then divide these counts by the total instances of the respective UMI across all observed guide assignments within that cell. For UMI-guide assignment combinations where this fraction is less than 20%, we do not count the UMI toward the final observed guide assignment counts. While this has some impact on the raw data, we find the benefits to be modest.

To make a set of final assignments, we take all whitelist sequences that have over ten reads and account for over 7.5% of the whitelist reads assigned to a given cell, where multiple sequences can be assigned to each cell. This set of assignments is merged with the filtered gene expression matrices output by cellranger such that only assignments to the filtered cells appear in the final data set.

Note that when processing CROP-seq data without PCR enrichment, we lowered the requirement for reads supporting a given guide to 3 to account for the decreased coverage of these transcripts.

2.3.15 Estimation of multiplicity of infection and capture rate

The most likely multiplicity of infection (MOI) and capture rate given the distribution of guide counts per cell were estimated using the generative model described in (Dixit et al., 2016). Briefly, a log likelihood is calculated using a zero-truncated poisson (MOI postselection) convolved with a binomial (incomplete capture of barcoded transcripts). This model is used to estimate the most likely set of MOI and capture rate values.

2.3.16 Monocle2 usage

PCA + tSNE, density peak clustering, differential expression testing, and size-factor estimation were performed using the monocle2 (Qiu et al., 2017a) functions reduceDimension, clusterCells, differentialGeneTest, and estimateSizeFactors unless otherwise noted.

2.3.17 Removing low-quality cells

We consistently observed a cluster of cells with much lower UMI counts on average than the rest of the data set when performing dimensionality reduction. To avoid including these cells in downstream analysis, we perform a simple procedure to remove any cluster with low average UMI counts. We perform PCA followed by TSNE on genes expressed in at least 50 cells for each condition, perform density peak clustering on two-dimensional tSNE space, calculate the average size factor over each cluster, and filter out clusters of cells with an average size factor of $2^{-}0.85$ or lower before downstream analysis.

2.3.18 Simulating loss in power from barcode swapping

Assignments were permuted for a fraction of cells ranging from 0 to 100% and kept fixed for the remaining fraction of cells. We tested for genes differentially expressed across the target assigned to each cell (testing genes detectably expressed in at least 50 cells; full model }textasciitilde target_gene). Differentially expressed genes at FDR of 5% were counted. Ten samplings were performed for each swap rate.

For the simulation performed on our own data, cells with a single target assignment from 100 nM doxorubicin-treated cells in our CROP-seq experiment were taken as the starting set of cells.

For the simulation on data from Adamson et al. (2016), processed data were obtained from GEO (GSE90546). Assignments of cells to targets were used as provided on GEO, and only cells noted as having high-quality assignment to a single target were used. Because of the large number of cells (50,000+) in the UPR experiment from this study and the large number of differential tests required for these simulations, the number of cells assigned to each target was downsampled two-fold to reduce runtime. We also performed tests on a data set further downsampled to approximately 6,000 cells to illustrate the impact of initial power.

2.3.19 tSNE embedding demonstrating TP53-enriched cluster

20 dimensions from PCA were carried into tSNE to two dimensions. All cells, including cells with guides to multiple targets and no assigned target, were included in dimensionality reduction for this plot. Percentages of cells with guides to TP53 and ARID1B were calculated, including cells that contain guides to multiple targets. All cells with TP53 guides were counted as TP53 cells only.

2.3.20 Enrichment of tumor suppressors in specific molecular states

Only cells containing a guide to a single target were considered in enrichment testing. A Chi-squared test was used to determine whether the distribution of individual sgRNAs and targets in tSNE space was significantly different from nontargeting controls at 5% FDR. Targets which did not pass this test and did not have an individual sgRNA pass the test were excluded from the subsequent enrichment tests. For each sgRNA of the remaining targets, we sought to estimate the functional editing rate (probability of a cell having a true LoF given that it received that sgRNA). Such estimates would be confounded if accounting for the possibility of edits that cause LoF for the target gene but have incomplete penetrance on the cellular phenotype. Therefore, we used an expectation-maximization approach to estimate the functional edit rate of each sgRNA relative to the unknown functional edit rate of the most efficient sgRNA for a given target.

The t-SNE cluster distribution of all cells in which a given sgRNA was detected was modeled as a mixture of the t-SNE cluster distribution of cells with a functional edit for the sgRNA's target gene and the t-SNE cluster distribution of nontargeting controls, where the mixing parameter is the relative functional edit rate for that sgRNA. In the expectation step, the t-SNE cluster distribution of cells with a functional edit for the target is estimated as the weighted average of the empirical t-SNE cluster distributions of each sgRNA for the target, weighted by the current estimates of the relative functional edit rate of the sgRNAs. In the maximization step, the relative functional edit rate of each sgRNA for the target is chosen to maximize the likelihood of the observed t-SNE cluster distribution for cells receiving that sgRNA under the multinomial mixture model.

After estimating the relative functional edit rate for each sgRNA, a weighted contingency table

was constructed where the rows are targets, the columns are t-SNE clusters, and the values are weighted cell counts, and where a cell's weight is proportional to the relative functional edit rate for the sgRNA it received. Fractional values were rounded down. Fisher's exact test was applied to this weighted contingency table to test for enrichment of targets amongst t-SNE clusters. Targets were defined as enriched at an FDR of 10%. Chi-square and Fisher's exact test were performed using R functions chisq.test and fisher.test, respectively.

2.3.21 Principal component and gene set enrichment analysis

Pairwise differential gene expression analysis was performed between enriched target cells and nontargeting controls for cells in all significantly enriched target–cluster pairs from our enrichment testing. The union of all differentially expressed genes across targets (FDR 5%) was used to perform principal component analysis. Gene set enrichment analysis was performed on genes that had the highest positive and negative loadings for principal component 1 (less than –0.02 or greater than 0.02). Gene set enrichment analysis was performed using the piano R package and the hallmarks gene set from MSigDB. Gene sets were defined as enriched at an FDR of 1%. PCA was performed using the prcomp function in R.

2.3.22 Code availability

Code and information on how to access additional data files relevant for secondary analysis can be found on Github at `https://github.com/shendurelab/single-cell-ko-screens`.

2.3.23 Data availability

Data is available on GEO via accession GSE108699 and is also provided via the Github repository described in "Code availability." pHAGE-GFP, pHAGE-BFP, and the CROP-seq vector with the CRISPRi-optimized backbone sequence described in the Online Methods are available on Addgene as 106281, 106282, and 106280. All CROP-seq experiments, except for the one presented in Figure 2d, were carried out with the original CROP-seq vector described in Datlinger et al. (2017). For

the experiments shown in Figure 2d, we used our own version of CROP-seq modified to contain a backbone optimized for CRISPRi, available on Addgene as described above.

2.3.24 Project acknowledgments

We thank all members of the Shendure and Trapnell labs for feedback on our manuscript and helpful discussions, particularly S. Srivatsan, G. Findlay, A. McKenna, R. Daza, B. Martin, M. Kircher, D. Cusanovich, X. Qiu, and V. Ramani. We thank J. Bloom and D. Fowler for discussions about lentivirus, and K. Han, J. Ousey, and M. Bassik for experimental advice and reagents for CRISPRi experiments. A.J.H. thanks Stella the cat for support. This work was supported by the following funding: NIH DP1HG007811 and UM1HG009408 (to J.S.), DP2HD088158 (to C.T.), and the W.M. Keck Foundation (to C.T. and J.S.). A.J.H. and M.J.G. are funded by the National Science Foundation Graduate Research Fellowship. J.L.M. is supported by the NIH Genome Training Grant (5T32HG000035) and the Cardiovascular Research Training Grant (4T32HL007828). C.T. is partly supported by an Alfred P. Sloan Foundation Research Fellowship. J.S. is an Investigator of the Howard Hughes Medical Institute.

Chapter 3: A GENOME-WIDE FRAMEWORK FOR MAPPING GENE REGULATION VIA CELLULAR GENETIC SCREENS

Chapter 3 is adapted with minimal modification from:

Gasperini, M.J., Hill, A.J., McFaline-Figueroa, J.L., Martin, B., Kim, S., Zhang, M.D., Jackson, D., Leith, A., Schreiber, J., Noble, W.S., Trapnell, C., Ahituv, N., Shendure, J. (2019) A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. Cell. *176*, 377–390.

3.1 ABSTRACT

Over one million candidate regulatory elements have been identified across the human genome, but nearly all are unvalidated and their target genes uncertain. Approaches based on human genetics are limited in scope to common variants and in resolution by linkage disequilibrium. We present a multiplex, expression quantitative trait locus (eQTL)-inspired framework for mapping enhancer-gene pairs by introducing random combinations of CRISPR/Cas9-mediated perturbations to each of many cells, followed by single-cell RNA sequencing (RNA-seq). Across two experiments, we used dCas9-KRAB to perturb 5,920 candidate enhancers with no strong a priori hypothesis as to their target gene(s), measuring effects by profiling 254,974 single-cell transcriptomes. We identified 664 (470 high-confidence) cis enhancer-gene pairs, which were enriched for specific transcription factors, non-housekeeping status, and genomic and 3D conformational proximity to their target genes. This framework will facilitate the large-scale mapping of enhancer-gene regulatory interactions, a critical yet largely uncharted component of the cis-regulatory landscape of the human genome.

3.2 INTRODUCTION

Consequent to an era of biochemical surveys of the human genome (e.g., Encyclopedia of DNA Elements [ENCODE]) and "common variant" human genetics (i.e., genome-wide association study

[GWAS] and expression quantitative trait locus [eQTL] studies), we are awash in candidate regulatory elements and phenotype-linked haplotypes, respectively (ENCODE Project Consortium, 2012; MacArthur et al., 2017). Determining whether and how each candidate regulatory element is truly functional, as well as pinpointing which variant(s) are causal for each genetic association, will require functional characterization of vast numbers of sequences.

We and others have recently adapted cell-based CRISPR/Cas9 genetic screens to evaluate candidate regulatory sequences in their native genomic context (Canver et al., 2015; Diao et al., 2016, 2017; Fulco et al., 2016; Gasperini et al., 2017; Klann et al., 2017; Korkmaz et al., 2016; Rajagopal et al., 2016; Sanjana et al., 2016). However, two aspects of these studies limit their scalability. First, they focus on the regulation of a single gene per experiment, typically entailing the development of a gene-specific assay. Second, each cell is a vehicle for one CRISPR-mediated perturbation, with the specificity-conferring guide-RNAs (gRNAs) usually introduced via lentivirus at a low multiplicity of infection (MOI). With millions of candidate regulatory elements and ˜20,000 regulated genes in the human genome, these limitations preclude the comprehensive dissection of the cis-regulatory architecture of even a single cell line.

Here, we introduce a framework (Figure 3.1A) designed to overcome both limitations. First, by using single-cell RNA sequencing (scRNA-seq) instead of gene-specific assays, one experiment can globally capture perturbations to gene expression (Adamson et al., 2016; Datlinger et al., 2017; Dixit et al., 2016; Hill et al., 2018; Jaitin et al., 2016; Xie et al., 2017), with no strong a priori hypothesis as to the target gene of each regulatory element tested. Second, by introducing gRNAs at a high MOI, each individual cell acquires a unique combination of perturbations against the isogenic background of a cell line. Introducing multiple perturbations per cell markedly increases power (Figure 3.1B). An association framework inspired by eQTL studies (Morley et al., 2004; Stranger et al., 2012) is used to map cis and trans effects by comparing gene expression in the subset of cells that contain a given gRNA to those that lack that guide. This strategy is analogous to conventional eQTL studies, but with individuals replaced by cells, variants replaced by unique combinations of gRNAs per cell to induce multiplex CRISPR-interference (CRISPRi), and tissue-level RNA-seq replaced by scRNA-seq. However, unlike eQTL studies, the resolution of our screen

is not constrained by linkage disequilibrium, nor is it limited to studying sites in which common genetic variants happen to exist. Although we recognize the imperfection of the analogy given that a reverse genetic screen using CRISPRi is far from equivalent to mapping the natural genetic variation that underlies QTLs, the fact that we were directly inspired by the eQTL framework led us to originally term this method “crisprQTL mapping.”

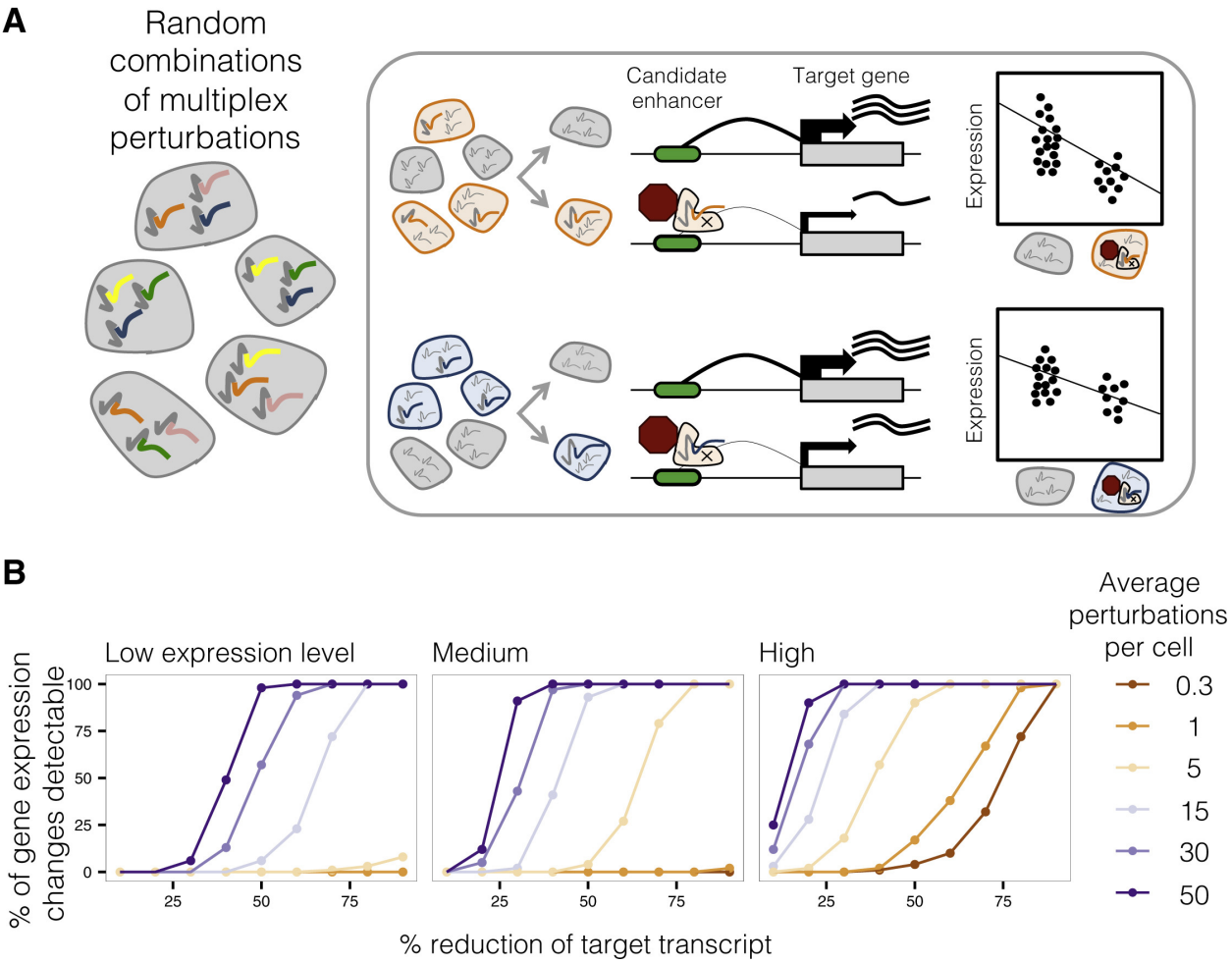


Figure 3.1: Multiplex Enhancer-Gene Pair Screening. A) Enhancer-gene pairs are screened by introducing random combinations of CRISPR/Cas9 candidate enhancer perturbations to each of many cells, followed by scRNA-seq to capture expression levels of all transcripts. Then, all candidate enhancers are tested against any gene by correlating presence of any perturbation with reduction of any transcript. B) Multiplex perturbations increase power to detect changes in expression in single-cell genetic screens while greatly reducing the number of cells that need to be profiled. Power calculations on simulated data show that increasing the number of perturbations per cell increases power to detect changes in expression, including for genes with low (0.10 mean UMIs per cell), medium (0.32), or high (1.00) mean expression. x axis corresponds to the simulated % repression of target transcript.

3.3 RESULTS

3.3.1 A Proof-of-Concept Multiplex Enhancer-Gene Pair Screen Targeting 1,119 Candidate Enhancers

To establish the feasibility of the assay formerly known as crisprQTL mapping, we targeted 1,119 candidate enhancers in the chronic myelogenous leukemia cell line K562, with CRISPRi as our mode of perturbation. For CRISPRi, we used a nuclease-inactive Cas9 tethered to the KRAB repressor domain to induce heterochromatin across an ˜1–2 kb window around a gRNA's target site (Thakore et al., 2015). The 1,119 candidate enhancers were all intergenic DNase I hypersensitive sites (DHSs) representing various combinations of H3K27 acetylation, p300, GATA1, and RNA Pol II binding (Figure 3.2A). Candidate enhancers were required to fall within the same toplogically associated domain (TAD) as at least one gene from the top decile of K562 expression and were collectively distributed across 510 TADs on every chromosome (Rao et al., 2014). 5,611 of the 12,984 genes expressed in K562 cells fell within 1 Mb of at least one candidate enhancer (K562-expressed genes defined as those observed in at least 0.525% of cells profiled in this pilot experiment).

Two gRNAs were designed to target each candidate enhancer. Additional pairs of gRNAs served as positive controls (targeting the transcription start sites [TSSs] of genes sampled from the top decile of K562 expression, or alternatively hypersensitivity sites of the α-globin locus control region [LCR]) and negative controls (50 non-targeting controls or "NTC" that target nowhere or in a gene desert).

This gRNA library was cloned into the lentiviral CROP-seq vector modified to include a CRISPRi-optimized backbone (Chen et al., 2013; Datlinger et al., 2017; Hill et al., 2018), and K562 cells were transduced at a high MOI (Figure 3.2B). After 10 days to allow for effective CRISPRi, the transcriptomes of 47,650 single cells were profiled. With a targeted amplification protocol (Adamson et al., 2016; Dixit et al., 2016; Hill et al., 2018), we identified a median of 15 ± 11.3 gRNAs per cell (Figure 3.2C). Each candidate enhancer or control was targeted in a median of 516 ± 177 cells (Figure 3.2D). For each targeted element, we partitioned the 47,650 cells based

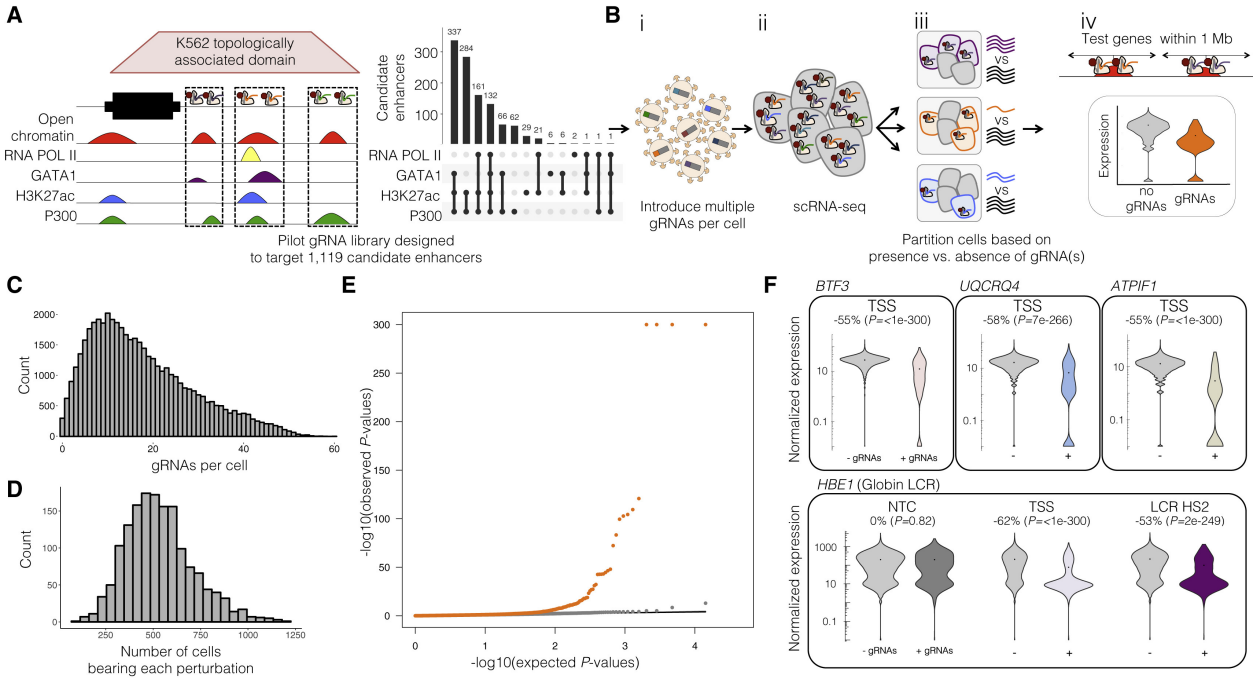


Figure 3.2: Pilot Multiplex Enhancer-Gene Pair Screen Testing 1,119 Candidate Enhancers in K562 Cells. A) 1,119 candidate enhancers were chosen based on intersection of enhancer-associated features and each targeted by two gRNAs. B) Schematic of this multiplex enhancer-gene pair screening method. (i) gRNAs were cloned into a lentiviral vector, and delivered to K562 cells at a high MOI. (ii) scRNA-seq was performed on these cells, with concurrent capture of the multiple gRNAs present in each cell. (iii) For each candidate enhancer, cells were partitioned based on whether or not they contained a gRNA targeting it. (iv) For each such partition, we tested for differential expression between the two populations for any gene within 1 Mb of the candidate enhancer. C) gRNAs were delivered to K562 cells at a high MOI, with median of 15 $\pm$ 11.3 gRNAs identified per cell. D) A total of 47,650 single cell transcriptional profiles were generated. Each perturbation was identified in a median of 516 $\pm$ 177 cells. E) Quantile-quantile plot of the differential expression tests. Distributions of observed versus expected p values for candidate enhancer-targeting gRNAs (orange) and NTC gRNAs (gray; downsampled) are shown. F) Expression of selected TSS (top row) and β-globin LCR positive controls (bottom row). Nearly all targeted TSSs, and all positive controls, showed significant differential expression of the expected target genes between cells with (+) versus without (-) targeting gRNAs, in contrast with NTCs. Percent changes and p values show the effect size and significance of differential expression of the denoted target gene between these cell groups.

on whether they did or did not contain gRNA(s) targeting it. We then tested for a reduction in the expression of each K562-expressed gene within 1 Mb of that element (Figure 3.2B) (Stranger et al., 2012). We also tested the 50 NTCs against all K562-expressed genes within 1 Mb of any targeted candidate enhancer. For perspective, with a "one gRNA per cell" framework, achieving equivalent power would require profiling the transcriptomes of ˜715,000 single cells.

A quantile-quantile plot showed an excess of significant associations involving the targeting of candidate enhancers relative to NTC controls (Figure 3.2E). We defined a 3.5% empirical false discovery rate (FDR) threshold based on the NTC tests as they are subject to the same sources of error as the element-targeting gRNAs. At this threshold, 94% (357 of 381) of TSS-targeting positive controls repressed their associated genes, as did all β-globin LCR controls (examples shown in Figure 3.2F). Additionally, we re-identified a known enhancer 3.6 kb upstream of GATA1 (Fulco et al., 2016).

At this same threshold, targeting of 11% of candidate enhancers (128 of the 1,119) repressed 1+ gene(s) within 1 Mb. As there were 13 candidate enhancers whose targeting impacted more than one gene (Figure 3.3A), this analysis yielded a total of 145 enhancer-gene pairs. Of the 105 downregulated target genes (Figure 3.3B), 26 were impacted by targeting of more than one of the 128 candidate enhancers (Figure 3.3A).

We examined the characteristics of paired enhancers whose targeting significantly impacted expression of 1+ genes in cis. We found paired candidate enhancers to be enriched for high chromatin immunoprecipitation sequencing (ChIP-seq) peak strength (based on average enrichment in ChIP-seq peak region) for enhancer-associated histone modifications (H3K27ac, logistic regression p value = 4e-5, candidate enhancers in the top quintile were 1.4-fold more likely to be paired than those in the bottom quintile), certain co-activators (p300, p value = 4e-16, 1.1-fold) and lineage-specific transcription factors (TFs) (GATA1 p value = 2e-7, 1.4-fold; GATA2 p value = 3e-10, 1.5-fold; SMAD1 p value = 1e-6, 1.4-fold; TAL1 p value = 6e-6, 1.1-fold; CCNT2 p value = 3e-7, 1.4-fold), whereas RNA Pol II and H3K4me1 were not associated (Figure 3.3C). Using these features, as well as average enrichment within the DHS and whether each had been previously validated in vivo (Visel et al., 2007), we trained a multivariate logistic regression classifier to

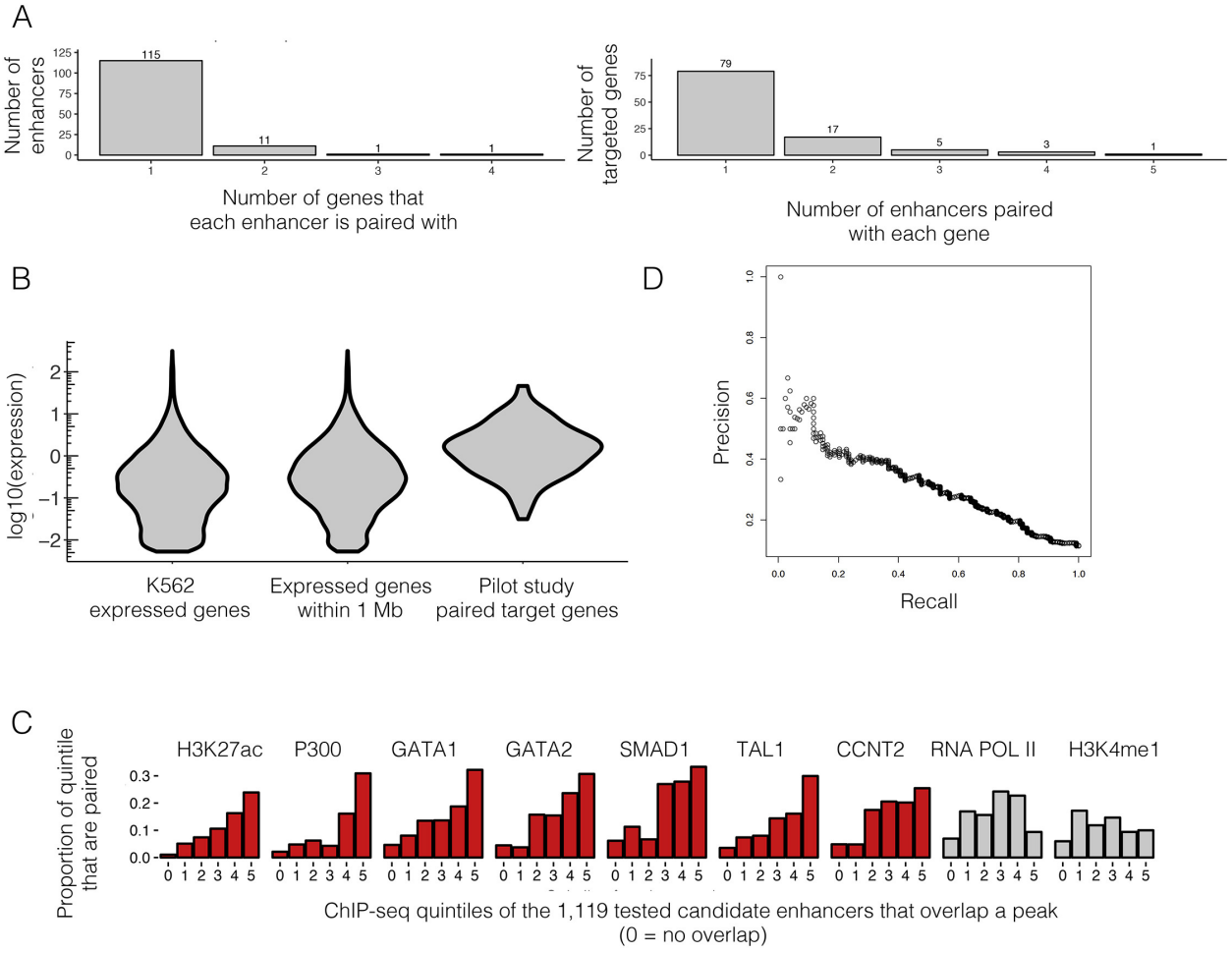


Figure 3.3: Details of 145 Enhancer-Gene Pairs Originally Identified in the Pilot Screen. A) Histogram per enhancer of the number of genes paired with that enhancer (3.5% empirical FDR in pilot screen, left). Histogram per gene of the number of enhancers paired with that gene (3.5% empirical FDR in pilot screen, right). B) Expression of target genes paired with candidate enhancers in the pilot screen. expression = mean transcript UMIs/cell in the entire 47,650 cell pilot dataset for: K562 expressed genes; those that fell within 1 Mb of a targeted candidate enhancer in the pilot experiment; and for the 105 genes targeted in the pilot experiment's pairs. In the pilot screen, tested candidate enhancers were required to fall within TADs that contained genes highly expressed in K562s. As these were then only tested for pairing with genes within 1 Mb, the pilot screen's target genes are potentially biased toward being highly expressed. This enrichment for highly expressed genes is not seen in the at-scale experiment, where tested candidate enhancers were not required to be in the same TAD a highly expressed gene (Figure 3.10C). (*legend continued on next page*)

(*continued*) C) Relative to the 1,119 candidate enhancers tested, the 128 paired candidate enhancers from the pilot experiment (3.5% empirical FDR) tend to fall in enhancer-associated ChIP-seq peaks that show stronger signals. All ChIP-seq peaks that overlap the 1,119 candidate enhancers were divided into quintiles of strength, defined as the average enrichment in ChIP-seq peak region (0 = no such peak overlaps the candidate enhancer, 1 = lowest, 5 = highest). Histograms of the proportion of each 1,119-quintile that were called as enhancer-gene pairs are shown. Red coloring = P-value ¡ 0.005 for independent logistic regression for predicting a candidate enhancer as paired based on this peak type. D) Precision-recall curve for a multivariate logistic regression classifier based on ENCODE enhancer-associated biochemical features that differentiates the 128 paired candidate enhancers from the remaining of the 1,119 candidate enhancers. The median AUPR from five-fold cross-validation was 0.31.

distinguish the 128 paired candidate enhancers from the 991 candidate enhancers for which we did not identify a target gene, achieving an AUPR of 0.31 (area under precision-recall curve; median from 5-fold cross validation; Figure 3.3D).

3.3.2 A Scaled Multiplex Enhancer-Gene Pair Screen Targeting 5,779 Candidate Enhancers

To demonstrate this approach at a substantially greater scale, we performed a second experiment targeting five times as many candidate enhancers (n = 5,779). First, two-thirds of these (n = 3,853) were new DHSs chosen by the classifier trained on the first experiment (Figures 3.4A and 3.3D). Second, as this set may be biased toward annotations used to select the initially targeted candidate enhancers (Figure 2A), we also targeted 948 exploratory DHSs chosen independent of the model (see STAR Methods). Third, we re-targeted 978 of the 1,119 initially targeted pilot candidate enhancers, including the aforementioned candidate enhancers paired with target genes in the pilot. Altogether, candidate enhancers targeted in this scaled experiment were within 1 Mb of 10,560 of 13,135 K562-expressed genes. As previously, we designed two gRNAs per candidate enhancer.

However, to evaluate whether poorly efficacious gRNAs might contribute to false negatives, we designed an additional two gRNAs for 377 of the 978 re-targeted candidate enhancers (Figure 3.4B). Finally, in addition to gRNA pairs targeting 5,779 candidate enhancers, we included the same positive and negative control gRNA pairs targeting 381 TSSs, the globin LCR, and 50 NTC pairs.

K562 cells were transduced at an even higher MOI than in the proof-of-concept experiment. We profiled the transcriptomes of 207,324 single cells and identified a median of 28 ± 15.3 gRNAs per cell (Figure 3.4C). Each candidate enhancer was targeted in a median of 915 ± 280 single cells (Figure 3.4D). Testing for associations as previously, a quantile-quantile plot again showed an inflation of significant associations involving the targeting of candidate enhancers (Figure 3.4E). Using the NTCs to set a more inclusive empirical FDR of 10%, 97% (369 of 381) of TSS-targeting positive controls repressed their associated genes, as did all β-globin LCR controls. At this same threshold, of the 5,779 candidate enhancers, we identified 600 as repressing 1+ gene(s) within 1 Mb. These included 397/3,853 model-selected candidate enhancers (10%), 35/948 systematically sampled exploratory DHS (4%), and 168/978 previously targeted candidate enhancers (17%). As targeting of 53/600 candidate enhancers downregulated more than one gene (Figure 3.4F), we collectively identified a total of 664 enhancer-gene pairs. As 113 genes were downregulated by targeting of more than one candidate enhancer, these pairs involved 479 target genes (Figure 3.4G). These ranged in effect size from -1.4% to -97.5% target gene repression (Figure 3.4H).

To evaluate reproducibility, we compared our results for the 978 candidate enhancers targeted in both experiments. Applying the same empirical FDR threshold of 10% to each dataset, 187/978 were identified as paired candidate enhancers in the pilot experiment, and 168/978 as paired candidate enhancers in the scaled experiment. Of these, 105 were identified in both experiments (hypergeometric test of overrepresentation p value 7e-45; 3.3-fold enriched over expectation). The pairs identified in both experiments had stronger effect sizes (median 25% versus 13% repression), better correlated effect sizes (Spearman's rho for % repression: 0.82 versus 0.13; Figure 3.5A), and involved more highly expressed genes (median 0.90 versus 0.63 UMIs per cell), than pairs identified in only one experiment.

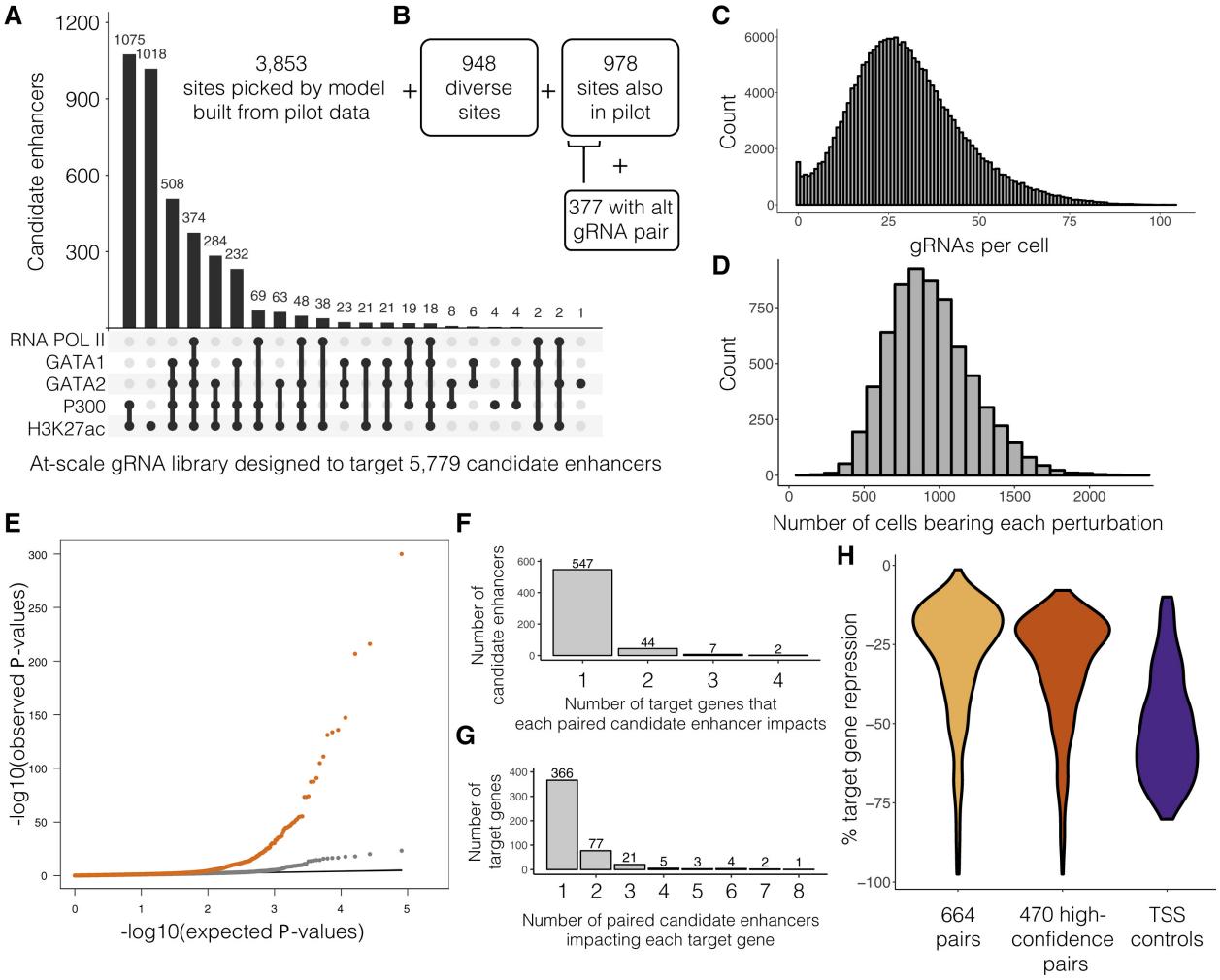


Figure 3.4: Multiplex Enhancer-Gene Pair Screening at Scale in K562 Cells. A) For a scaled experiment, gRNAs were designed to target a total of 5,779 candidate enhancers. Characteristics are shown for 3,853 sites chosen by a model informed by the hits identified in the pilot experiment. B) 948 exploratory candidate enhancers were sampled from K562 DHSs. 978 candidate enhancers from the pilot were re-targeted with the same gRNA pair, and 377 of these were also targeted with a second, alternative gRNA pair. C) gRNAs were again delivered to K562 cells, but at a higher MOI than the pilot experiment (median 28 ± 15.3 gRNAs identified per cell). D) A total of 207,324 single cell transcriptional profiles were generated. Each perturbation was identified in a median of 915 ± 280 single cells. (*legend continued on next page*)

(*continued*) E) Q-Q plot of the differential expression tests. Distributions of observed versus expected p values for candidate enhancer-targeting gRNAs that were correlated with decrease in target gene expression (orange) and NTC gRNAs (gray; downsampled) are shown. F) Histogram of the number of target genes impacted by each candidate enhancer identified as part of a pair (10% empirical FDR). G) Histogram of the number of paired candidate enhancers detected as regulating each target gene (10% empirical FDR). H) Effect sizes for the 664 enhancer-gene pairs that pass a $<$0.1 empirical FDR, the 470 high-confidence enhancer-gene pairs, and the 97% of TSS controls that are detected as repressing their target genes.

As noted above, an additional pair of gRNAs for 377/978 re-targeted candidate enhancers were included in this experiment, to facilitate evaluation of the extent to which poorly efficacious gRNAs might contribute to false negatives. In the scaled experiment at a 10% empirical FDR, 109/377 of the original gRNA pairs and 119/377 of the new gRNA pairs mediated enhancer-gene pairs. Of these, 84 were directed at the same candidate enhancers, a highly significant overlap (hypergeometric test of overrepresentation p value 4e-33; 2.4-fold enriched over expectation). Furthermore, the effect sizes on the most highly repressed genes for gRNA pairs targeting the same candidate enhancer were well-correlated (Spearman's rho for % repression: 0.73; Figure 3.5B). Overall, this analysis suggests that targeting candidate enhancers with more than two gRNAs could modestly increase our sensitivity.

Due to the noise from variability in expression levels, effect sizes, and gRNA quality, we defined a high-confidence subset of reproducible enhancer-gene pairs as those identified in both experiments at the 10% empirical FDR (112 pairs; 359/381 [94%] of targeted TSSs also met this criteria), as well as those internally reproducible between the 2 independently targeting gRNAs for candidate enhancers only tested in the scaled experiment (358 pairs; 337/381 [88%] of targeted TSSs also met this criteria). Putting these sets together, we annotated 470 enhancer-gene pairs as high-confidence, involving 441 candidate enhancers (Figure 3.5C) and impacting expression of

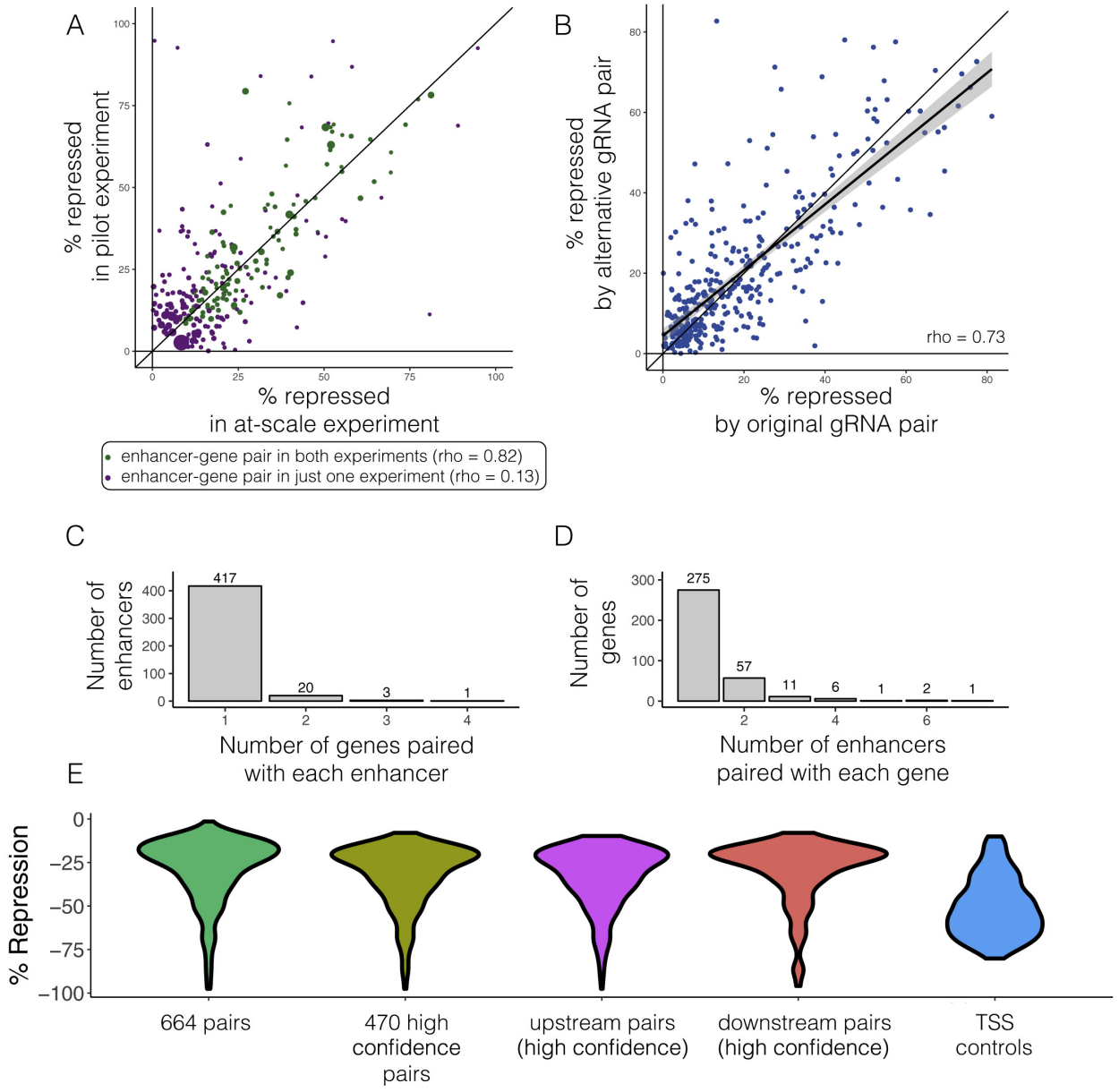


Figure 3.5: Replication of Effect across Experiments and Alternative gRNA Pairs. A) The percent target gene repression of an enhancer-gene pair in the pilot versus the scaled experiments (green: called as a pair in both experiments; purple: called as pair in only one experiment). B) The effect sizes on the most highly repressed gene for each pair of gRNA pairs targeting the same candidate enhancer (see STAR Methods). C) Histogram per enhancer of the number of genes paired with that enhancer (high confidence pairs of the at-scale screen). (*legend continued on next page*)

(*continued*) D) Histogram per gene of the number of paired with that gene (high confidence pairs of the at-scale screen). E) Effect sizes from enhancer-gene pairs identified in the at-scale screen. % repression of target transcript for the 664 enhancer-gene pairs that pass a < 0.1 empirical FDR, the 470 high confidence enhancer-gene pairs, the high confidence pairs in which the enhancer is upstream of the target gene, high confidence pairs in which the enhancer is downstream, and the 97% of 381 TSS controls that are detected as repressing their target genes in the at-scale screen.

353 target genes. These ranged in effect size from -7.9% to -97.5% (Figure 3.4H). We use this high-confidence subgroup for all summary analyses described below, unless otherwise noted. Of note, 24 candidate enhancers are paired with multiple target genes (Figure 3.5D); it is possible that some of these pairings represent indirect effects (e.g., if a gene that is the primary target of the enhancer is involved in the regulation of the other gene).

3.3.3 Replication or Validation of 22 Selected Enhancer-Gene Pairs in Singleton Experiments

We next sought to individually replicate 15 enhancer-gene pairs with a range of effect sizes (-10% to -81%) and 6 "null" candidate enhancers not paired with any target gene. We transduced K562 cells separately with small pools of gRNAs targeting individual candidate enhancers, and investigated the impact on gene expression via bulk RNA-seq. For 12/15 replication experiments targeting candidate enhancers associated with downregulation of a target gene, the effect sizes were similar in magnitude and direction of effect (Figures 3.6A–D and 3.11). For all 9 experiments predicted to cause $>$30% repression, replication effects were also significant in a test of differential expression (cis adjusted p value $<$0.1). Of the 6 lines targeting a "null" candidate enhancer, none significantly decreased expression of a gene located within 1 Mb of the target (cis adjusted p values $>$0.1).

Although the field often refers to singleton independent re-testing via CRISPRi as "validation," it is a recapitulation of the modality of perturbation of the screen and perhaps better classified as

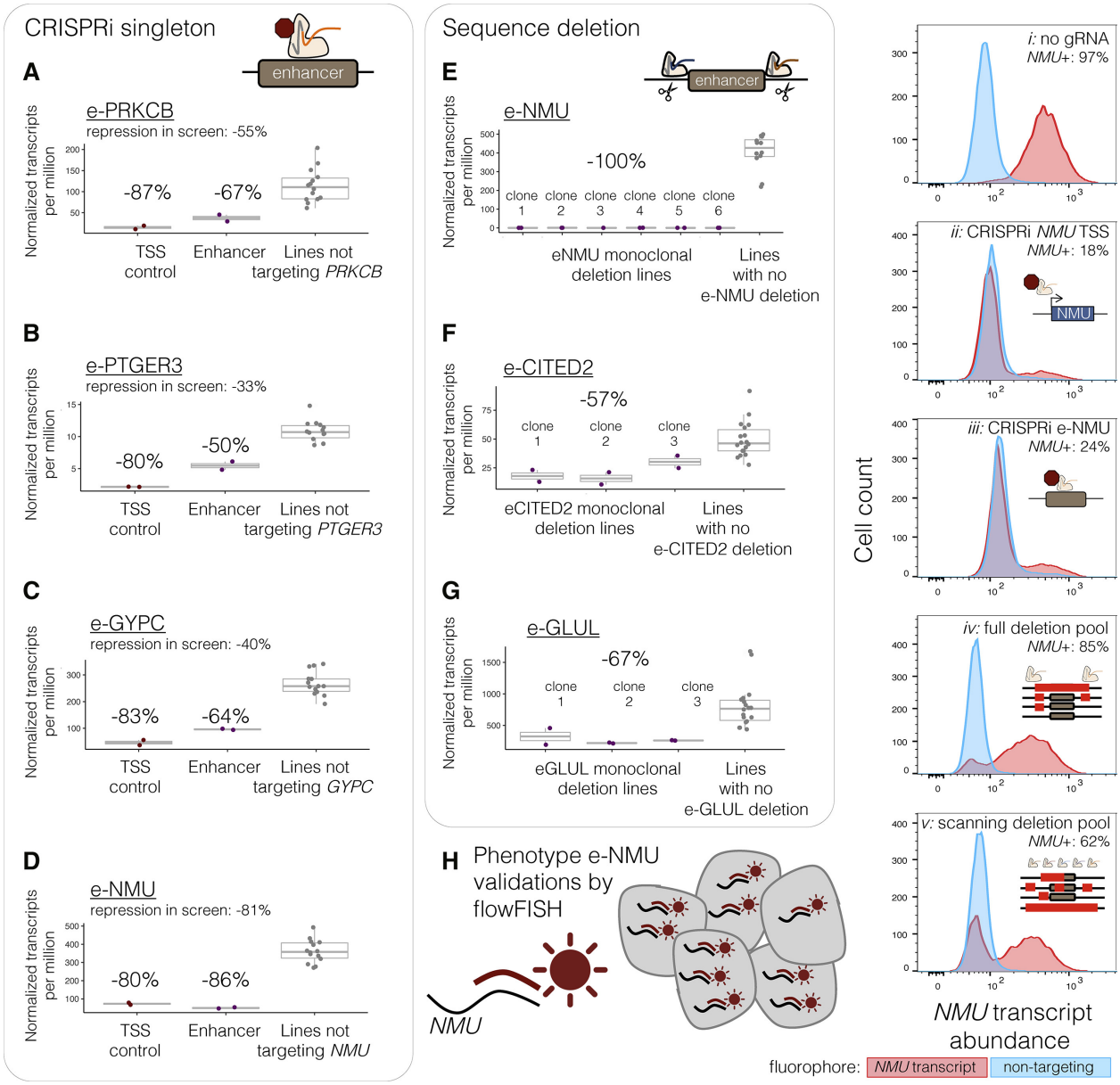


Figure 3.6: Replication and Validation of Selected Enhancer-Gene Pairs in Singleton Experiments. A–D) For each singleton replication experiments of enhancer-gene pairs, bulk RNA-seq was performed on CRISPRi+ K562 cells transduced with gRNAs targeting (purple) e-PRKCB (A), e-PTGER3 (B), e-GYPC (C), e-NMU (D), or the TSSs (dark red) of their respective target genes. Target gene expression in the singleton-target cell lines (red/purple) as compared to replication experiments in which the other 4 candidate enhancers or TSSs were targeted (gray). Eleven other singleton CRISPRi experiments are summarized in Figure 3.11. (*legend continued on next page*)

(*continued*) E–G) To validate three enhancer-gene pairs by sequence deletion, monoclonal lines were generated with full deletion of the locus's genomic sequence in three to six independent clones (e-NMU, E; e-CITED2, F; and e-GLUL, G), followed by bulk RNA-seq. See also Figure 3.8A. H) NMU-targeting cells were phenotyped by fluorophore-labeling of intracellular NMU transcripts by RNA flowFISH. (ii–iii) Singleton CRISPRi targeted cells as in (D). (iv–v) A heterogeneous pool of cells engineered such that a portion (based on deletion efficiency) harbor full or scanning deletions of e-NMU.

another form of replication. Therefore, we also performed a more stringent validation by generating 3+ monoclonal homozygous deletion lines for each of 3 enhancers (effect size in scRNA-seq screen: e-NMU = -81%, e-CITED2 = -35%, e-GLUL = -21%; Figure 3.8). All three selected enhancers are quite distal from the gene whose expression they regulate (>50 kb). These homozygously deleted lines all had the expected and magnitude of direction of effect (Figure 3.6E–G), indeed with stronger effect sizes than seen by CRISPRi perturbation in the scRNA-seq screen (effect size with deletion: e-NMU = -100%, e-CITED2 = -57%, e-GLUL = -67%).

In our validations of the NMU candidate enhancer ("e-NMU"), we also applied RNA flowFISH (Choi et al., 2018) and again observed decreased NMU expression in singleton CRISPRi populations targeting NMU's TSS (-79% less NMU than untreated cells) and e-NMU (-73% less NMU, Figure 3.6H, ii–iii). We also used flowFISH to phenotype a heterogeneous pool of cells that harbored a mix of full, partial, or no deletions of e-NMU, generated by transient transfection of flanking pairs of gRNAs. 12% of the cells showed reduced NMU expression in comparison to untreated cells (Figure 3.6H, iv), which is in-line with expected full deletion efficiency (**?**). Cells were sorted into bins of low, medium, or high NMU expression. PCR of the e-NMU locus revealed enrichment of the full deletion in the low and medium NMU bins, whereas full deletion was rarer in the high NMU bin (Figure 3.8B). To further dissect e-NMU, we additionally transfected with 19 gRNAs interspersed every ˜100 bp across e-NMU to generate deletions of diverse lengths and

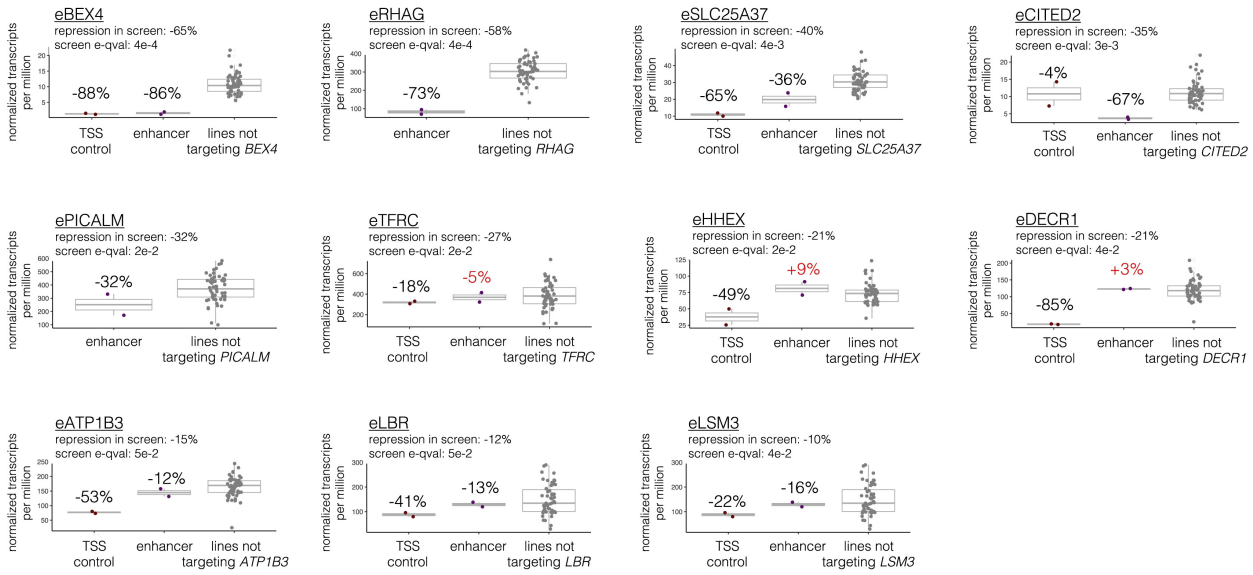
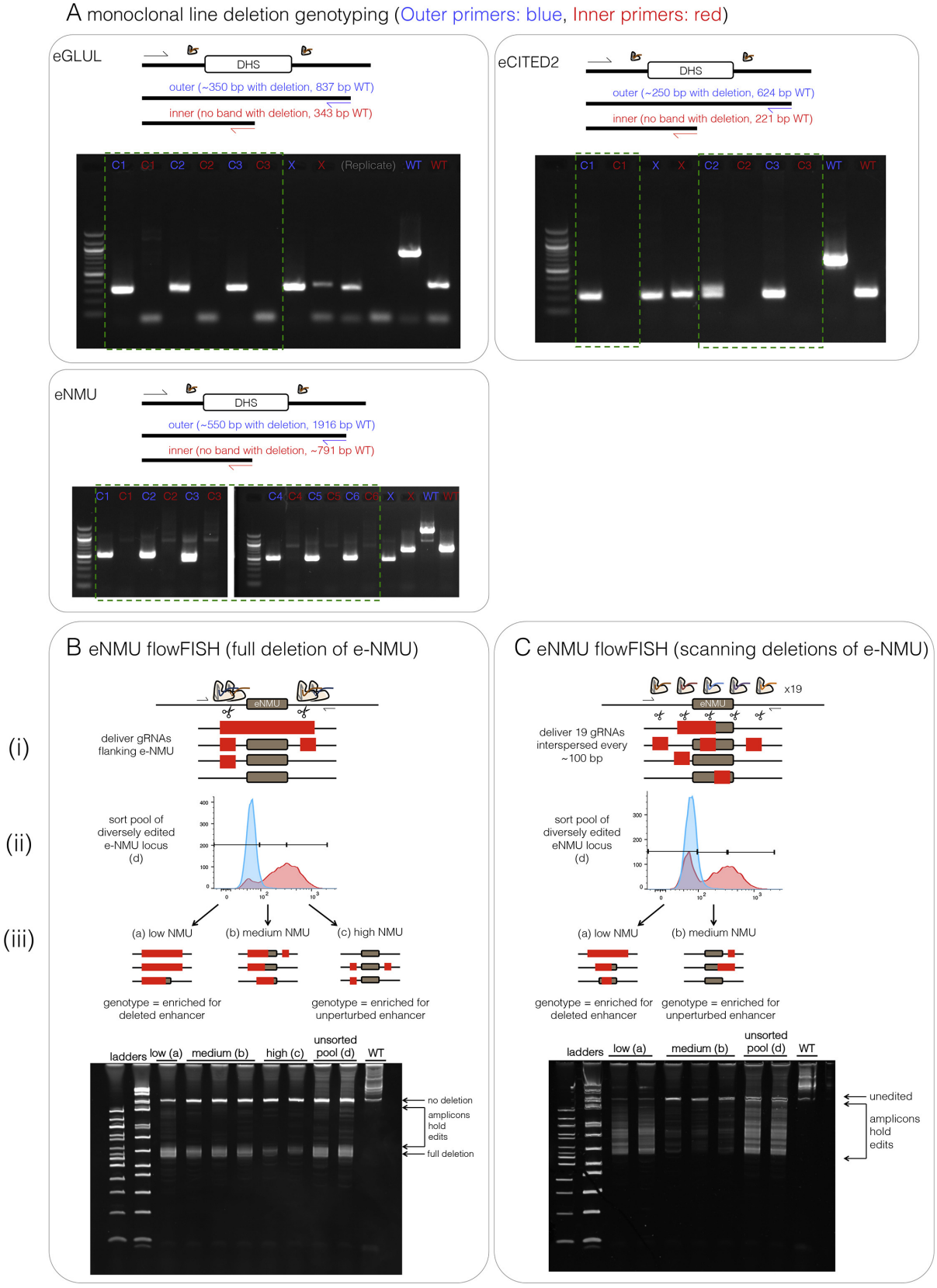


Figure 3.7: Eleven Further Singleton CRISPRi Experiments. For each singleton replication experiments of enhancer-gene pairs, bulk RNA-seq was performed on CRISPRi-positive K562 cells transduced with CRISPRi-optimized CROP-seq gRNAs targeting the labeled paired-enhancer (purple, denoted with an "e" prefix) or the TSSs (dark red) of their respective target genes. Target gene transcript expression in the singleton-target cell lines (dark red/purple) as compared to 'non-targeting' lines (gray; singleton experiments in which the other 10 candidate enhancers or TSSs were targeted plus a line transduced with non-targeting gRNAs). Repression in screen = differential expression from at-scale screen. Screen e-qval = Benjamini-Hochberg corrected empirical P-value from at-scale screen. Normalized transcripts per million (tpm) from sleuth. %s above boxplots = sample's % repression in bulkRNA-seq calculated from (transcript's mean tpm between the sample's two technical replicates) / (transcript's mean tpm from all the 'non-targeting' lines). % repression labeled light red if in disagreement with the enhancer-gene pair in the at-scale screen.



(*legend on next page*)

Figure 3.8: Details of Sequence Deletion Validation. A) Genotyping PCR design and gels for the homozygous sequence deletion monoclonal lines, as featured in Figures 3.6E–G. Outer primers were designed to amplify the entire candidate enhancer locus; shorter band in these 'outer' lanes (blue label, as compared to 'WT' lane) represents presence of a full deletion. Inner primers were designed to amplify only if a wild-type allele remained ("red" labeled lanes); presence of a band indicates a remaining wild-type locus. Primers design is schematized at the top. Clones with a deletion band (in the "outer" PCR lane) and no wild-type band (in the "inner" PCR lane) were submitted to bulkRNA-seq. Green dashed outline represents the clones used in Figure 3.6. Nomenclature of "C1" and "C2" etc correspond to "clone 1" and "clone 2" et cetera as labeled in Figure 3.6. 'WT' lanes = same parental K562 cell line that was transfected with gRNA targeting HPRT1. Ladder = NEB 100 bp (N3231L). "X" = cell line did not harbor homozygous deletions. B and C) e-NMU sequence-disrupted cells were phenotyped by NMU RNA flowFISH, as featured in Figure 3.6H. First, K562 cells were transfected with nuclease-active Cas9 and gRNAs either flanking (B) or scanning (C) the e-NMU locus to create a heterogeneous population of cells (i) in which a portion (based on editing efficiency) harbor full or partial deletion of e-NMU. Then, intracellular NMU expression was labeled via flowFISH (ii) and cells were sorted into bins of low (a), medium (b), or high (c) NMU expression (as to sort genotypes based on the effect upon disruption of e-NMU function, iii). Last, gDNA was extracted from the cells in each bin, and the e-NMU locus was amplified (primers diagrammed at the top of the figure). Unsorted pool (d) = unsorted-but-edited cells to demonstrate original distribution of genotypes in the original heterogenous pool. Each lane is a replicate PCR of gDNA (10 ng per reaction) from that same sorted sample. Ladder L = 100 bp (NEB), Ladder R = 1 Kb ext (Invitrogen). WT = untreated parental K562s. Remaining full-length alleles in the 'low' expression bins could correspond to inaccuracy of flowFISH, alleles with very small edits, or (as K562s are pseudotriploid) heterozygous cells that still retained a largely uninterrupted copy of e-NMU on one or two alleles.

locations, inducing reduction of NMU expression in 35% of cells compared to untreated (Figure 3.6H, v). PCR of e-NMU again showed a similar enrichment of longer deletions in the cells with lower NMU expression (Figure 3.8C).

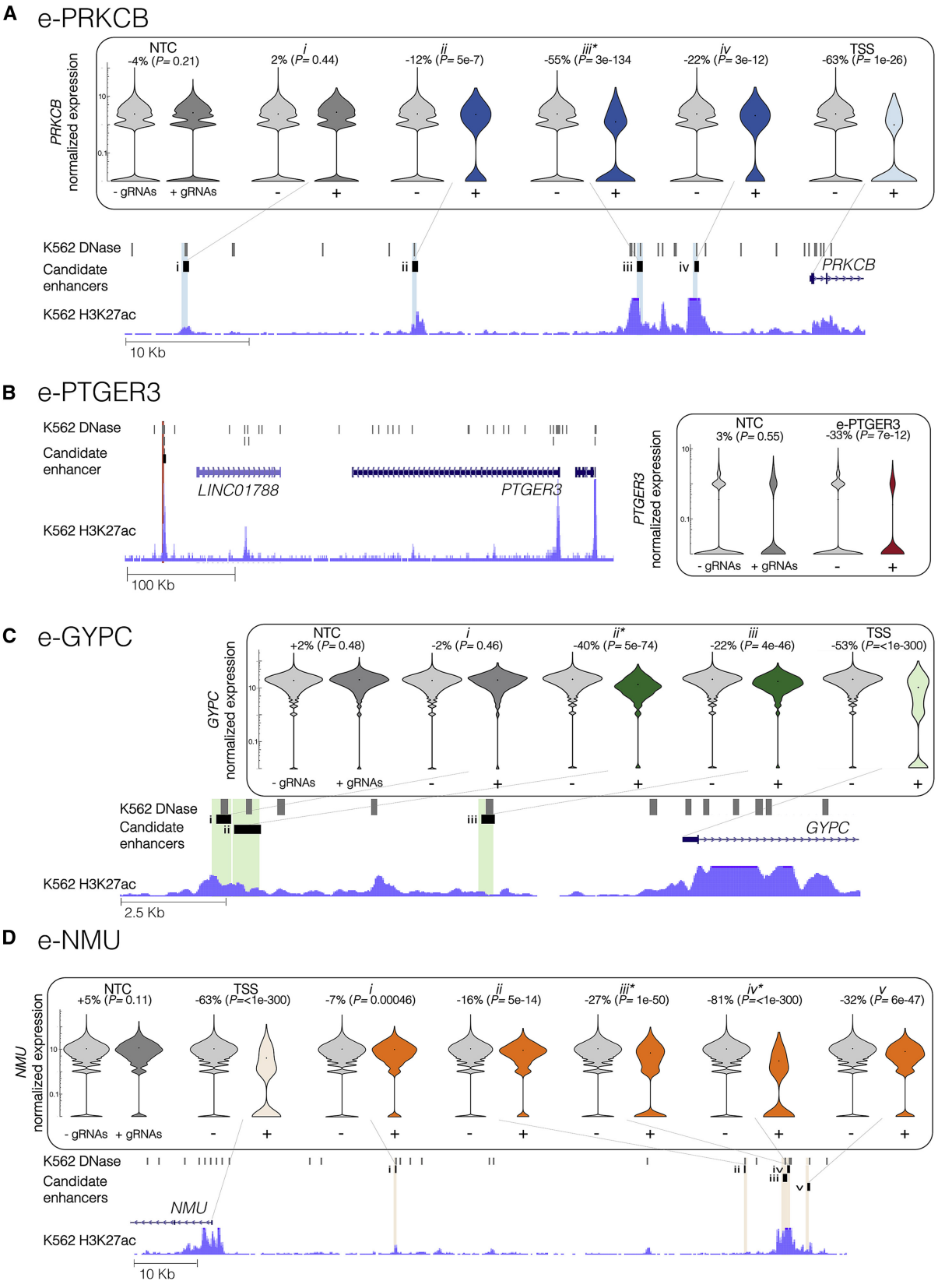
In summary, of the high confidence pairs that we re-tested by singleton CRISPRi and/or singleton CRISPR-mediated deletion, 13/16 matched with respect to both their direction and magnitude of effect size, whereas 3/16 failed to validate. This false-positive rate is consistent with the 10% FDR that we used to assign a threshold for calling pairs (p value on whether 3/16 disagrees with 10% FDR = 0.21).

3.3.4 Selected Examples of Enhancer-Gene Pairs

We highlight four of the enhancer-gene loci in Figure 3.9. An "e-" prefix is used to denote candidate enhancers that we targeted in singleton replication experiments. In the scaled experiment, we targeted four candidate enhancers across the region upstream of PRKCB. The furthest of these (Figure 3.9A, i; 50 kb upstream) did not have an effect, but candidate enhancers 32, 14, and 9 kb upstream of the TSS were associated with repression of PRKCB (Figure 3.9A, ii–iv). The strongest of these, located 14 kb upstream, was also targeted and replicated in both the pilot and singleton experiments ("e-PRKCB", Figure 3.6A and Figure 3.9A, iii).

In the pilot, scaled, and singleton replication experiments, we targeted only one candidate enhancer within 1 Mb of PTGER3 ("e-PTGER3," Figure 3.6B and Figure 3.9B), located 371 kb downstream of the PTGER3 TSS. In each of the three experiments, targeting of e-PTGER3 consistently repressed expression of PTGER3.

We targeted three candidate enhancers in the region upstream of GYPC, a human erythrocyte membrane protein. Targeting of candidate enhancers 4.5 kb upstream (Figure 3.9C, iii) and 10 kb ("e-GYPC", Figure 3.6C and Figure 3.9C, ii) upstream of GYPC's TSS resulted in its repression in the scaled experiment. Interestingly, a candidate enhancer so close to e-GYPC as to likely be unresolvable from it by CRISPRi (Figure 3.9C, i) did not result in repression of GYPC in the scaled experiment, potentially attributable to poor gRNA quality or another source of false negatives.



(*legend on next page*)

Figure 3.9: Highlighted Examples of Enhancer-Gene Pairs. A) Three candidate enhancers (labeled ii–iv) that reside 32, 14, and 9 kb upstream of PRKCB were paired with PRKCB, but a fourth (i) that lies 50 kb upstream was not (shown: hg19 chr16:23791225-23851797; iii is e-PRKCB in Figure 3.6A). B) A single candidate enhancer (e-PTGER3 in Figure 3.6B) located 371 kb downstream of PTGER3 was paired with PTGER3 (shown: chr1:71104684-71582921). C) Two candidate enhancers paired with GYPC (ii–iii) lie in the 11 kb region upstream of GYPC. However, a third candidate enhancer (i) immediately adjacent to (ii) was not paired with GYPC (shown: chr1:71104684-71582921; ii is e-GYPC in Figure 3.6C). D) Targeting five candidate enhancers (i–v) located 30.5, 87, 93.4, 94.1, and 97.6 kb upstream of NMU, significantly reduced expression of NMU (shown: chr1:71104684-71582921; iii-iv is e-NMU in Figure 3.6D). Target genes' normalized expression presented on log scale. Asterisks denote the candidate enhancers that were targeted as part of a singleton replication experiment (Figure 3.6). + and - denote the cells from the at-scale screen with or without gRNAs targeting that locus. Percent changes and p values denote the size and significance of a differential expression between these cell groups.

Targeting of multiple candidate enhancers decreased expression of the same gene, NMU, which encodes neuromedin U, a neuropeptide that plays roles in inflammation as well as erythropoiesis (Gambone et al., 2011). One candidate enhancer was associated with light repression of NMU (Figure 3.9D, i; located 30.5 kb upstream of the NMU TSS). An additional four candidate enhancers were located in close proximity to one another, but nearly 100 kb upstream of the NMU TSS ("e-NMU", Figure 3.6D and Figure 3.9D, ii–v; located 87, 93.4, 94.1, and 97.6 kb upstream). Because of their proximity, these closely located candidate enhancers internally replicate e-NMU within the scaled experiment, in contrast to the neighboring candidate enhancers of e-GYPC.

3.3.5 Distance between Paired Enhancers and Promoters

We find that of the class of enhancers surveyed here (non-intronic, unbuffered by other enhancers), paired enhancers are separated from the TSS of their target genes by a median distance of 24.1 kb (Figure 3.10A, top row). Note that this analysis is restricted only to high-confidence pairs that fall upstream of their target genes (n = 354), to avoid bias from the length of the gene body consequent to the fact that we avoided targeting intronic candidate enhancers for which CRISPRi might directly inhibit transcription. Upstream and downstream enhancers do not exhibit large differences in their effect size distributions (Figure 3.5E). Given that we tested for associations against all genes within 1 Mb of each candidate enhancer (Figure 3.10A, fourth row; median distance of 440.2 kb, similarly restricted to upstream tests), this supports a very strong role for proximity in governing enhancer-promoter choice. Nonetheless, 153/470 (33%) of enhancer-gene pairs involved skipping of at least one closely located TSS of another K562-expressed gene (Figure 3.10B). Interestingly, low-confidence enhancer-gene pairs (i.e., the subset of the 600 that were not high-confidence and also fall upstream; n = 127) were also enriched for proximity to their target genes, suggesting that a substantial proportion of these are bona fide enhancers (Figure 3.10A, second row; median distance of 45.0 kb).

Of our 359 "positive control" TSSs whose targeting successfully repressed the expected gene in both experiments, 35 reduced expression of 1+ additional genes (45 apparent promoter-promoter relationships in total). 15 of these 45 involved overlapping promoters (TSSs within 1 kb), such that the observed effect of CRISPRi is likely direct. As for the remaining 30, one possibility is that these represent examples of promoters acting as enhancers, as recently reported (Diao et al., 2017; Fulco et al., 2016). Additionally, as repressive epigenetic effects may spread a few kilobases from the target site, it is possible that CRISPRi of promoters may be silencing proximal enhancers as well. However, these 30 are largely not enriched for proximity to affected genes (Figure 3.11A; median distance of 405.3 kb, similarly restricted to upstream tests), in contrast with enhancer-gene pairs (median distance = 24.1 kb). We therefore hypothesize that these are more likely consequent to trans effects of repressing the primary target of these TSS-targeting gRNAs. In other words, rather

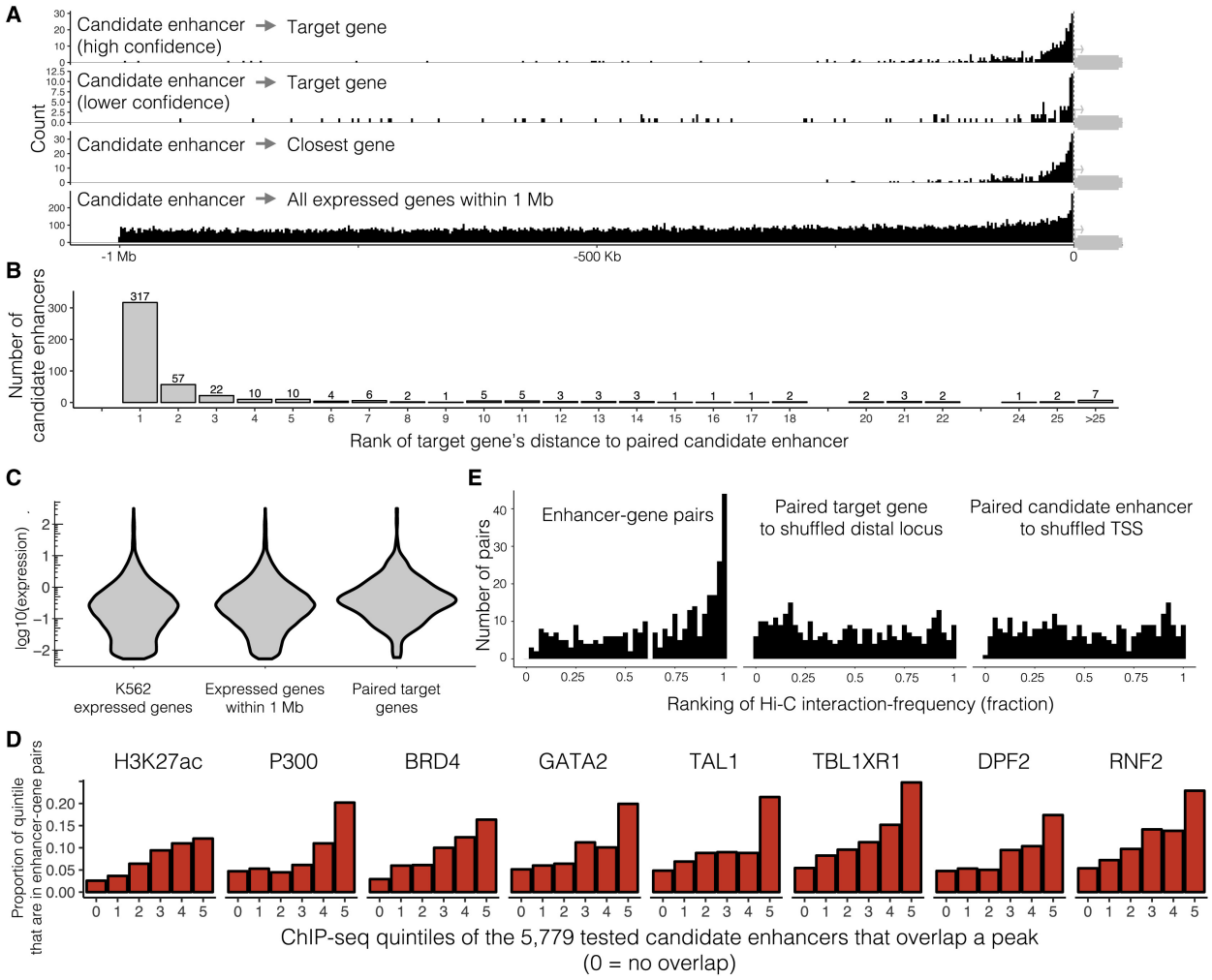


Figure 3.10: Characteristics of K562 Enhancer-Gene Pairs. A) Paired candidate enhancers fall close to target genes. Distribution of distances between the paired candidate enhancers and their target gene's TSS (top row, high confidence pairs; second row, lower confidence pairs), the TSS of whatever K562-expressed gene is closest (third row), or the TSS of every K562-expressed gene within 1 Mb (fourth row). Plotted with respect to gene orientation. Of the 470 high confidence pairs, this plot displays only the 354 that fall upstream of the target genes (as the gRNA library does not include candidate enhancers within 1 kb of any gene body, downstream enhancers are biased to fall further from the target TSS). A TSS-focused zoom of this plot is included as Figure 3.11E. B) 317 of 470 high-confidence pairs target the most proximal K562-expressed gene. Target genes are ranked by their absolute distance to the paired candidate enhancer (1 = closest, 2 = second closest, etc.). (*legend continued on next page*)

(*continued*) C) This framework captures regulatory effects on genes from a broad range of expression levels (expression = mean transcript UMIs/cell in the entire 207,324 cell dataset, for 13,135 K562-expressed genes, 10,560 of these within 1 Mb of a targeted candidate enhancer in the scaled experiment, and 470 high-confidence enhancer-gene pairs). See also Figure 3.11D. D) Paired candidate enhancers tend to fall in enhancer-associated ChIP-seq peaks that show stronger signals. All ChIP-seq peaks that overlap the scaled experiment's 5,779 candidate enhancers were divided into quintiles defined as the average enrichment in ChIP-seq peak region (0 = no such peak overlaps the candidate enhancer, 1 = lowest, 5 = highest). Histograms of the proportion of which candidate enhancers in each quintile that were paired with a target gene are shown for the eight most-enriched ChIP-seq datasets. E) Enhancer-gene pairs interact more frequently in K562 Hi-C data (left, fractional ranking of enhancer-gene pairs' Hi-C interaction-frequency against all other possible interactions at similar distances within the same TAD, K-S test against a uniform distribution p value ¡2e-16), as compared to two control distributions: paired target gene TSSs paired with a shuffled genomic locus (middle: K-S test versus actual enhancer-gene pairs distribution = p value 2e-7) or paired candidate enhancers paired with a shuffled genomic locus (right, K-S test versus actual enhancer-gene pairs distribution = p value 1e-9). See also Figures 3.11B and C.

than these gRNA-targeted promoters acting as noncoding regulatory elements of other genes, the reduction in protein levels of the targeted gene may secondarily affects the expression of other genes.

3.3.6 Characteristics of Target Genes

The 353 genes included in 1+ 470 high-confidence enhancer-gene pairs had several notable characteristics. First, their expression levels are distributed similarly to the full set of 10,560 genes against which we tested (Figure 3.10C), suggesting we are reasonably well-powered to detect regulatory effects on even modestly expressed genes. Second, housekeeping genes were underrep-

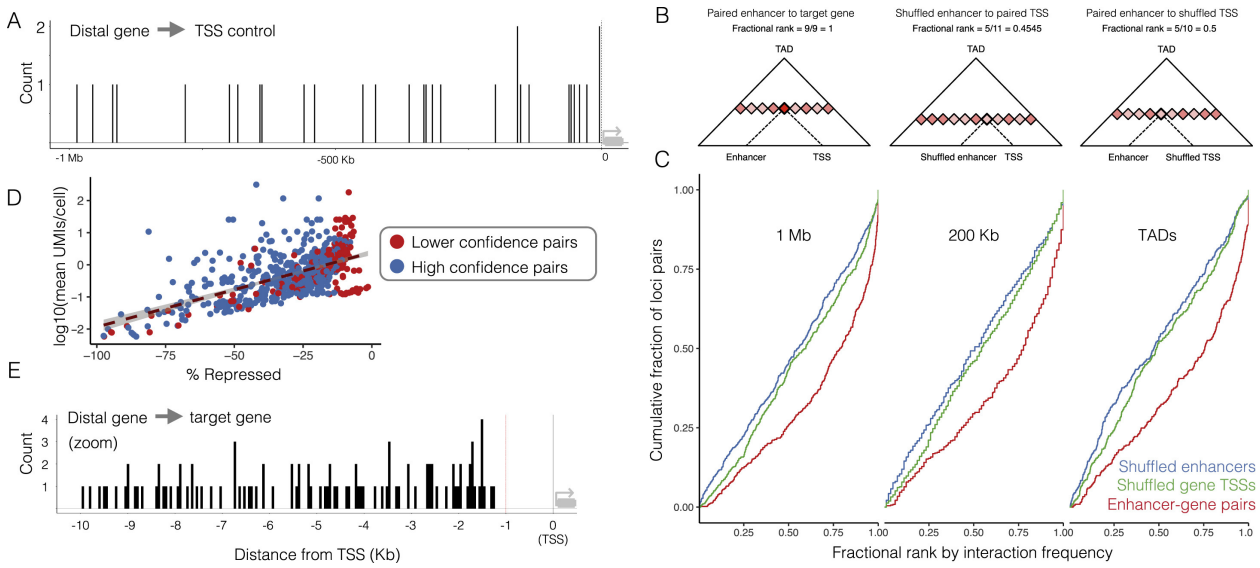


Figure 3.11: Details on Characteristics of K562 Enhancer-Gene Pairs. A) Distribution of distances between "positive control" TSSs and any secondarily repressed genes. Of our 359 'positive control' TSSs whose targeting successfully repressed the expected gene in both experiments, 35 reduced expression of 1+ additional genes (45 apparent promoter-promoter relationships in total). 15 of these 45 involved overlapping promoters (TSSs within 1 Kb) and are not shown here as the observed effect of CRISPRi is likely direct. The distances that the remaining 30 secondarily repressed genes fall upstream of the targeted TSS are shown. In contrast with enhancer-gene pairs (Figure 3.10A), these 30 are largely not enriched for proximity to affected genes. Dashed line = target gene TSS. B and C) Hi-C interaction frequency analysis, (B) Example schematic of fractional ranking by interaction frequency analysis. The interaction frequency of each loci pair (color of pixel) is ranked within the interaction frequencies of all distance-matched genomic-pairs in the same TAD (the stripe of pixels shown in schematic). For the two null distributions in Figure 3.10E, each pair's target gene's TSS is given a shuffled enhancer (and then ranked again within this new distance distribution), or the pair's candidate enhancer is given a shuffled TSS (and then ranked again within this new distance distribution). Shuffled TSSs and enhancers are drawn from the same distance distribution as the actual enhancer-gene pairs. (C) The same fractional rank by interaction frequency analysis within the same TAD as shown in Figure 3.10E, but also comparing ranking to all pairs within 1 Mb or 200 Kb of the chosen enhancer-TSS pair. Red = enhancer-gene pairs, blue = hit-gene to shuffled enhancer pair null distribution, green = hit enhancer to shuffled gene TSS pair null distribution. (*legend continued on next page*)

(*continued*) D) Correlation of effect size of enhancer-gene pair versus expression level of target gene. Effect size (% transcript repressed) was correlated with expression level of targeted gene (Spearman's rho for 664 inclusive pairs: 0.56; Spearman's rho for 470 high confidence pairs: 0.53). This is likely consequent to power, as small effects (less than -25%) are not detected on lowly expressed genes (less than 0.12 UMIs/cell). log10 of the mean UMIs/cell is denoted per target gene transcript. E) A "zoom-in" of Figure 3.10A to the 10 Kb upstream of the target gene's TSS (rather than 1 Mb). 101 of 354 upstream, high confidence enhancer-gene pairs fall within 10 Kb of the TSS. Same restrictions to enhancer-gene pairs plotted here as in Figure 3.10A. Gray line = TSS, red line = 1 Kb upstream of TSS (all protospacers within 1 Kb of a TSS were excluded from any candidate enhancer gRNA library).

resented, relative to all tested genes (hypergeometric test p value = 3e-5 and 2.1-fold depleted using the housekeeping gene list of (Eisenberg and Levanon, 2013); hypergeometric test p value = 2e-6 and 3.9-fold depleted using the housekeeping gene list of Lin et al. (2017)). Similar depletions of housekeeping genes are observed when we instead compare paired target genes to the K562-expressed genes most proximal to tested candidate enhancers. Although these analyses support the view that a prevailing characteristic of housekeeping genes may be a dearth of distal regulatory elements (Ganapathi et al., 2005; Gasperini et al., 2017), we cannot fully rule out that the possibility that this result is influenced by our choice of candidate enhancers to target. Finally, paired target genes were enriched for genes with roles in leukocyte migration and differentiation, consistent with distal enhancers shaping the expression of K562-specific genes.

3.3.7 Characteristics of Paired Enhancers

We also examined the characteristics of the candidate enhancers for which targeting significantly impacted expression of 1+ genes in cis. First, as compared with the full set of 5,779 candidate enhancers targeted in either or both experiments, we tested if the 441 high-confidence candidate

enhancers were enriched for strong peaks in 169 K562 ChIP-seq datasets (ENCODE Project Consortium, 2012). We identified 87 that were significantly enriched (threshold of an adjusted p value $<$0.005), but the eight most significant were co-activators (p300 logistic regression p value = 1e-46, candidate enhancers in the top quintile were 1.8-fold more likely to be paired than those in the bottom quintile; BRD4 p value = 2e-33, 1.6-fold), an enhancer-associated histone modification H3K27ac (p value = 8e-37, 1.6-fold), the MYC activator TBL1XR1 (p value = 2e-34, 1.5-fold), and lineage-specific TFs (TAL1 p value = 2e-33, 1.6-fold; GATA2 p value = 1e-31, 1.5-fold; DPF2 p value = 5e-31, 1.5-fold; RNF2 p value = 2e-33, 1.5-fold; Figure 3.10D). Other expected enhancer-associated marks also exhibited significant enrichment (CCNT2 p value = 4e-21, 1.3-fold; H3K4me1 p value = 1e-19, 1.8-fold; MYC p value = 2e-12, 1.3-fold). However, many of these features are correlated, and BRD4, H3K4me1, TRIM24, p300, H3K27ac, ETS1, and ZNF274 were the only significant predictors in a multivariate logistic regression (p value $<$ 0.01). Of note, high conservation as measured by median phyloP scores (Pollard et al., 2010) was not enriched in these candidate enhancers as compared to all tested candidate enhancers (independent logistic regression p values $>$ 0.5).

Second, we examined whether paired enhancers were more likely to intersect with K562 super-enhancers. Overall, 474 of the 5,779 candidate enhancers that we tested fell within 65 K562 super-enhancers (Cao et al., 2017); however, a much higher proportion of high-confidence paired enhancers belonged to this set (102/441). Several super-enhancers contained multiple targeted enhancers that were paired with the same gene. More specifically, 20 genes were linked with two candidate enhancers, and 6 genes were linked with three or four candidate enhancers, that were located within the same super-enhancer.

Third, we evaluated enrichment of TF motifs in either our associated enhancers or the promoters of their target genes. Motifs for the known blood TFs KLF-1, -5, -6, -15, leukemogenesis-related SALL4, and the MYC-interacting ZN281 were enriched in the promoters of the inclusive set of 479 paired-target genes, as compared to the promoters of all genes within 1 Mb of a tested candidate enhancer. Similarly, motifs for a largely distinct set of known blood TFs (TAL1, KLF-1, -3, -4, -5, -8, and GATA-1, -2, -3) and AP2C were enriched in the inclusive set of 600 paired

enhancers, as compared to the overall set of 5,779 candidate enhancers tested.

3.3.8 Pairs of Transcription Factors Act Together across Enhancer-Gene Pairs

To investigate whether there was any discernible logic underlying why particular enhancers were associated with particular promoters, we next sought to identify pairs of TFs that are "co-enriched" in the inclusive set of 664 enhancer-promoter pairs (i.e., they occur across pairs at a higher frequency than expected by chance given their background frequency in each category). We identified 6 TF pairs whose sequence motifs were co-enriched in this way, suggesting potential interactions. For example, presence of the NR2C2 motif (implicated in regulation of the globins (Tanabe et al., 2007)) in a paired promoter was associated with presence of a KLF1 or RXRA motif in the corresponding paired enhancer. On the other hand, presence of the GATA3 motif in a paired promoter was associated with the absence of a KLF1 motif in the corresponding paired enhancer.

We also explored such pairings via ChIP-seq data. Although ChIP-seq peaks often reflect indirect binding, such secondary partners might still play a role in the restriction of enhancer-promoter interactions. We identified 24 TF pairs that are "co-enriched" in enhancer-promoter pairs. Unfortunately, none of the TF pairs identified in either analysis had corresponding ChIP-seq datasets or high quality consensus motifs for both TFs involved in the pair, preventing cross-confirmation between the two modalities of analysis.

3.3.9 Comparison of Enhancer-Gene Pairs to Hi-C-Based Measurements of Physical Proximity

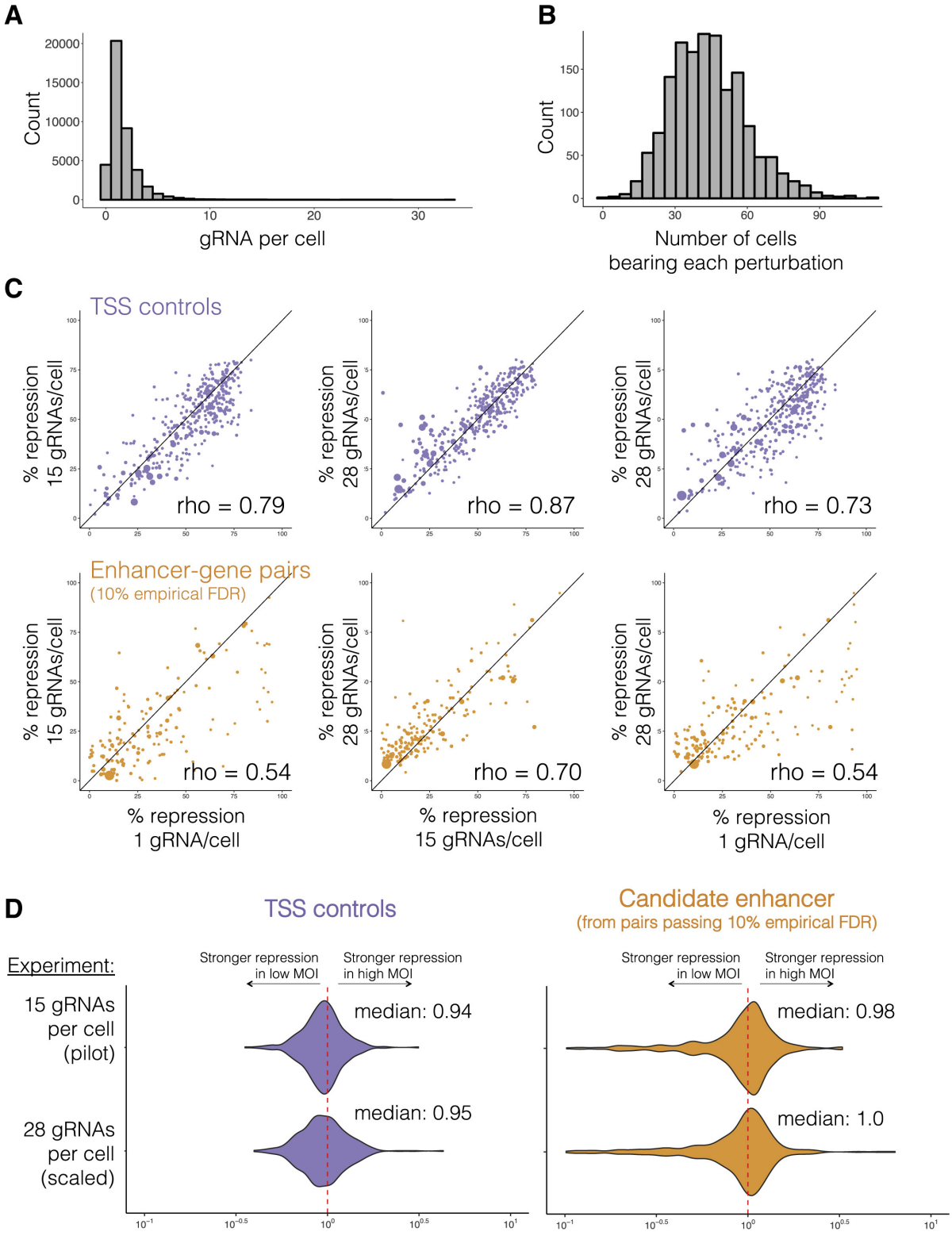
We sought to evaluate whether our enhancer-gene pairs are enriched for physical proximity as measured by the global chromosome conformation mapping technique Hi-C. To control for the dominant effects of genomic distance and TADs in Hi-C datasets, we ranked the Hi-C contact frequencies in K562 cells (Rao et al., 2014) for the 71% of the enhancer-gene pairs that fell in the same TAD (333/470 high-confidence pairs) against all other possible interactions at similar distances within the same TAD (median 66 other genomic-loci pairs, range 6 to 260, Figures 3.11B and C). Upon plotting the fractional ranks of high-confidence pairs, we found their contact

frequencies to be strongly enriched at the highest ranks (Kolmogorov-Smirnov [K-S] test against a uniform distribution p value <2e-16, Figure 3.10E). To ensure that this enrichment was not an artifact of paired enhancers or genes interacting more frequently with all neighboring loci (as in FIREs (Schmitt et al., 2016), we repeated this analysis twice but shuffled the genomic loci paired to either the enhancers or genes (keeping these shuffled pair sets' overall distance distributions the same as the original enhancer-gene pair set's distance distribution). This did not result in the same enrichment as seen in the high confidence pair distribution (K-S test of high confidence enhancer-gene pair versus enhancer-pair shuffling p value 1e-9; high confidence enhancer-gene pair versus TSS-pair shuffling p value 2e-7), consistent with more frequent looping specifically between the high confidence enhancer-gene pairs (Figure 3.10E). Although enriched for proximity, we note that only a minority of our hits are called as proximate to their target genes based on this analysis; as such, many enhancer-gene pairs would not have been identified if we had limited tested candidate enhancers to those physically proximate to a promoter according to Hi-C or related data.

3.3.10 CRISPRi Is Highly Multiplexable within Cells

To our knowledge, prior to this study, it was unknown whether extensively multiplexing gRNAs within a single cell would dilute the efficacy of CRISPRi. To evaluate this, we conducted a biological replicate of the pilot experiment, targeting the same 1,119 candidate enhancers but at a low MOI. From this experiment, we profiled the transcriptomes of 41,284 cells and identified a median of 1 $\pm$ 1.6 gRNAs per cell (Figure 3.12A). Each perturbation was only seen in a median of 43 $\pm$ 16 cells, as compared with 516 $\pm$ 177 cells in the high MOI pilot experiment (Figure 3.12B). At a 10% empirical FDR, only 316 TSSs and 69 enhancer-gene pairs were identified in the low MOI experiment, as compared with 359 TSSs and 226 enhancer-gene pairs in the high MOI pilot experiment, validating the substantial increase in power resulting from multiplexed perturbation (Figure 3.1B). As the same 381 TSS controls were targeted in the low MOI, pilot, and scaled experiments, we compared the degree of repression conferred by CRISPRi at increasing MOI (median 1 versus 15 versus 28 gRNAs per cell), and found them to be well-correlated (Spearman's rho's ranging

from 0.73 to 0.87; Figure 3.12C). On average, the degree of repression conferred by targeting a TSS in both high MOI experiments was only ~6% less than by targeting it in the low MOI experiment (Figure 3.12D). Similarly, for candidate enhancers paired in the scaled experiment (10% empirical FDR) that were also targeted in the low MOI and pilot experiments, effect sizes were well correlated (Spearman's rho's ranging from 0.54 to 0.70; Figure 3.12C), and effect sizes ratios clustered around 1 (Figure 3.12D). Overall, these results suggest that multiplexing gRNAs within individual cells, even to MOIs of ~28, does not dilute the efficacy of CRISPRi.



(*legend on next page*)

Figure 3.12: CRISPRi is Robust to Multiplexing within a Cell. (A) A biological replicate of the pilot study, targeting the same 1,119 candidate enhancers and 381 TSSs, was performed at a low MOI (median 1 $\pm$ 1.6 gRNAs identified per cell). B) A total of 41,284 single cell transcriptional profiles were generated. Each perturbation was identified in a median of 43 $\pm$ 16 single cells. C) Correlation of effect sizes for TSS controls (top, purple) or enhancer-gene pairs identified in the scaled experiment (10% empirical FDR, bottom, orange) across increasing rates of gRNA per cell (left, 1 versus 15; middle, 15 versus 28; right, 1 versus 28 gRNAs/cell). Point sizes are proportional to each target gene's expression level. D) The ratios of repression for each TSS control or paired candidate enhancer (as identified with a 10% empirical FDR in any experiment) in the low MOI experiment versus a high MOI experiment (top = median 1 gRNA versus 15 gRNAs; bottom = median 1 gRNA versus 28 gRNAs). The candidate enhancer outliers with stronger effect sizes in the low MOI experiment (right panel, ratios in long left tail) are likely largely due to stochastic undersampling of lowly expressed target genes in the low MOI experiment.

3.4 Discussion

Understanding the regulatory landscape of the human genome requires the validation and identification of target genes for the vast numbers of candidate enhancers that have been nominated by biochemical marks or that reside within haplotypes implicated by GWAS or eQTL studies. Our multiplexed enhancer-gene pair screening method has the potential to help address this challenge. In the scaled experiment, we evaluated 78,776 potential cis regulatory relationships involving 5,779 candidate enhancers and 10,560 expressed genes. In contrast, nine recently published CRISPR screens of noncoding sequences cumulatively studied regulatory effects on a total of 17 genes (Canver et al., 2015; Diao et al., 2016, 2017; Fulco et al., 2016; Gasperini et al., 2017; Klann et al., 2017; Korkmaz et al., 2016; Rajagopal et al., 2016; Sanjana et al., 2016). By delivering a median of 28 perturbations to each of 207,324 cells, this experiment was powered equivalently to a "one gRNA per cell" experiment profiling 5.8 million single cell transcriptomes. Of note, one recent

study used scRNA-seq as a readout for the effects of CRISPR-based perturbations of 71 candidate regulatory elements on ~100 genes in seven genomic regions (Xie et al., 2017). However, its power and scope was limited by a low MOI (Figure 3.1B) and a gRNA barcoding strategy that suffers from a ~50% rate of template switching (Hill et al., 2018; Xie et al., 2018).

For future iterations of target prioritization for multiplexed enhancer-gene pair screening, several characteristics of our identified enhancer-gene pairs are important to keep in mind. Foremost, although a wide range of effect sizes (7.9% to 97.5% for the 470 high-confidence pairs, Figure 3.4H) were observed on genes with a broad range of expression levels (0.0058 to 313 UMIs/cell, Figure 3.10C), effect sizes were correlated with expression levels (Spearman's rho 0.53; Figure 3.11D). This is likely consequent to power, as small effects are more challenging to detect on lowly expressed genes. Additionally, we note that although we identified many genomic features that were significantly correlated with the likelihood of belonging to an identified pair, a pilot-trained classifier informed by biochemical marks did not appreciably increase our hit rate in the at-scale screen. Furthermore (1) 29% of enhancers did not fall within the same TAD as their target gene, (2) although enriched for proximity in 3D space as measured by Hi-C, the majority of enhancer-gene pairs are not identified as contacts in such datasets, and (3) although enriched for sequence-level proximity, one-third of enhancer-gene pairs involved skipping of at least one closely located TSS of another K562-expressed gene. These observations underscore the difficulty of the prediction task, and we recommend that future screens do not overly bias themselves toward looking under the lamppost until additional examples accrue and the rules of mammalian gene regulation are better understood.

Although it may be surprising that cis changes in gene expression were identified for only ~10% of the candidate enhancers tested here, there are several potential caveats to bear in mind. First, previous studies have identified shadow enhancers acting to mask the effects of perturbing individual enhancers (Hong et al., 2008), although a genome-wide survey of such enhancer redundancy has yet to be conducted. To investigate such interactions more thoroughly, future iterations of our method could randomly distribute programmed pairs of multiplexed enhancer perturbations per locus. Second, other technical caveats include (1) not all enhancers may be susceptible to dCas9-

KRAB perturbation, (2) gRNAs may be variably effective in targeting enhancers (Figure 3.5B), (3) some enhancers required for the initial establishment rather than maintenance of gene expression could be missed in a screen in a stable immortalized cell line, and (4) we did not comprehensively survey the noncoding landscape surrounding each gene, and the marks we used to define candidate enhancers may be excluding some classes of distal regulatory elements. These caveats are respectively addressable in the future by using other epigenetic modifiers or nuclease-active Cas9, by using more gRNAs per candidate enhancer, by combinatorial perturbation of selected loci (Xie et al., 2017), by using cell models of differentiation, and by densely tiling selected loci with perturbations.

Nonetheless, the fact that our paired candidate enhancers are predicted by the strength of enhancer-associated marks (e.g., H3K27ac, p300) supports the assertion that we are identifying bona fide enhancers and simultaneously weakens the case for elements that were negative. Our study provides new insights into key properties of human enhancers, e.g., the distribution of distances between at least some types of enhancers (i.e., unbuffered, upstream) and their target genes. A full understanding of the precise rules governing enhancer-promoter choice is a topic of great interest and will be facilitated by the identification of more enhancer-gene pairs.

A limitation of enhancer-gene pair screening as implemented here relates to the resolution of CRISPRi. In the future, this can potentially be improved upon by adapting enhancer-gene pair screening to use single or pairs of gRNAs with nuclease-active Cas9 to disrupt or delete candidate enhancers at the sequence level. A separate concern is whether high MOI transduction is inducing a cellular inflammatory response, and therefore biasing discovery. However, although some genes with roles in inflammation are among our paired target genes (e.g., NMU, IL6), we only observed pathway-level enrichment of one immune-system related pathway. Moreover, the effect sizes observed in our high MOI versus low MOI experiments were well correlated.

To date, ENCODE has cataloged over 1.3 million human candidate regulatory elements based on biochemical marks (`http://screen.umassmed.edu/`), while GWAS have identified over 75,000 unique haplotype-trait associations (`https://www.ebi.ac.uk/gwas/`). Validating candidate elements, fine-mapping of causal regulatory variants, and identifying the target genes of both en-

hancers and regulatory variants, represent paramount challenges for the field. Given the scale of the problem, we anticipate that the multiplex, genome-wide framework presented here for mapping gene regulation can help overcome these challenges.

3.5 METHODS

3.5.1 Cell Lines and Culture

K562s cells are a pseudotriploid ENCODE Tier I erythroleukemia cell line derived from a female (age 53) with chronic myelogenous leukemia (Zhou et al., 2018). K562 cells expressing dCas9-BFP-KRAB (Addgene #46911, polyclonal) were a gift of the Bassik lab, grown at 37°C, and cultured in RPMI 1640 + L-Glutamine (GIBCO) supplemented with 10% fetal bovine serum (Rocky Mountain Biologicals) and 1% penicillin-streptomycin (GIBCO). K562s were authenticated by bulk/single-cell RNA-seq and visual inspection.

HEK293Ts (a human embryonic kidney female cell line) used for housemade virus production were cultured at 37°C in DMEM also supplemented with 10% fetal bovine serum and 1% penicillin-streptomycin. HEK293Ts were authenticated by visual inspection.

3.5.2 Note About Terminology Used Below

A gRNA-group is defined as all the gRNAs that are targeting the same candidate enhancer or positive control site. To note, all novel TSS and candidate enhancer targeting gRNA-groups are referred to as “perturbative gRNA-groups,” whereas all others are referred to as ”control gRNA-groups.”

3.5.3 Pilot Library - 1,119 candidate enhancers

Picking candidate enhancer regions: K562 DNase-seq narrowPeaks (ENCSR000EKS) ¡ 1 Kb away from any gene (GENCODE March 2017 v26lift37) were bedtools-intersected (Quinlan and Hall, 2010) with K562 Hi-C domains (Rao et al., 2014) that contained at least one of the top 10%

most highest expressed genes in a previously generated 6,806 single-cell K562 dataset. The remaining regions were largely taken from intersections with K562 GATA1 ChIP-seq narrowPeaks (ENCSR000EFT, lifted to hg19), H3K27ac ChIP-seq narrowPeaks (ENCSR000AKP, lifted to hg19), RNA Pol II ChIP-seq narrowPeaks (ENCSR000AKY), and EP300 ChIP-seq narrowPeaks (ENCSR000EHI) (Figure 3.2A). Ten further sites were handpicked and do not overlap either of these four marks.

Candidate enhancer gRNAs: NGG-protospacers within these candidate enhancers were scored using default parameters of FlashFry (McKenna and Shendure, 2018), and the two top-quality-scoring gRNA per region were chosen as spacers to be used in the gRNA library (scores prioritized by Doench2014OnTarget > Hsu2013 > Doench2016CDFScore > otCount).

TSS positive control gRNAs: 381 genes were randomly sampled from the highly-expressed genes within the same Hi-C domains (as described above) and 2 gRNA were chosen per gene from spacers with the best empirical and predicted scores of the hCRISPRiv2 library (Horlbeck et al., 2016). To note - these spacers are designed as 19 bp, rather than the full 20 of the spacers used in the rest of our gRNAs.

NTC gRNAs: 50 scrambled-sequence spacers with no targets in the genome and 11 protospacers targeting 6 gene-devoid regions of the genome (hg19 chr4:25697737-25700237, chr5:12539119-12541619, chr6:23837183-23839683, chr8:11072736-11075236, chr8:23768553-23771053, chr9:41022164-41024664) were chosen as evaluated by Benchling's CRISPR tool. These were randomly paired to create a gRNA group. More were chosen from 6 random regions of the hg19 genome (chr4:25697737-25700237, chr5:12539118-12541619, chr6:23837183-23839683, chr8:11072736-11075236, chr8:23768553-23771053, chr9:41022164-41024664) using FlashFry (McKenna and Shendure, 2018) to total 50 targeting these gene-devoid regions of the genome. A further 39 NTCs were sampled from those recommended by Horlbeck et al. (2016). A gRNA to the CAG promoter was additionally included as an internal control (labeled "cag_promoter" in relevant supplementary data, but excluded from analysis for simplicity).

Distal enhancer positive control gRNAs: 15 gRNAs targeting the HBE1 TSS, and HS1-4 of the Globin LCR were chosen as validated from Klann et al. (2017) and Xie et al. (2017). These were

manually paired based on their target sites to create gRNA-groups.

3.5.4 Note about Pilot Library

Our initial FlashFry quality annotations when designing the pilot experiment did not label a small number of protospacers with perfect repeat off-targets, permitting their inclusion in our library (81 of 2,238 spacers ordered in the pilot library; only 9 gRNA-groups with both spacers affected). gRNA-groups with an impacted spacer were rare in our 145 significant enhancer-gene pairs. We also note that we still expect these guides to target their intended site, but with potentially more off-targets. This error was fixed for evaluating gRNA quality in the scaled experiment.

3.5.5 At-Scale Library - 5,779 candidate enhancers

Choice of new and repeated sites: A logistic regression classifier built using the 145 enhancer-gene pairs originally identified in the pilot experiment (see Aggregate analysis of enhancer-gene pairs: ChIP-seq strength quintile analysis and logistic regression classifier) was used to select the top 5,000 intergenic open chromatin regions in K562s (as defined by DNase-seq narrowPeaks (ENCSR000EKS)). Of these, 3,853 were over 1 Kb away from boundaries (GENCODE March 2017 v26lift37) of any genes expressed in the pilot 47,650 K562 single-cell dataset, were not previously included in the pilot library, and had minimum two gRNAs with high quality as again determined by FlashFry. Of the top 5,000, 120 corresponded to a candidate enhancer in one of the original 145 pilot enhancer-gene pairs, and 851 of these corresponded to candidate enhancers targeted in the pilot library but not originally identified as part of a enhancer-gene pair. We additionally included 7 more candidate enhancers not top-ranked by our model, but identified as part of the original 145 enhancer-gene pairs. The only candidate enhancer that was identified in an original 145 pilot enhancer-gene pair but not included in this library had no high quality gRNAs by this second library's standards (see Note about Pilot Library). Only 15 sites did not overlap any of the marks shown in Figure 3.4A.

Two alternative gRNAs were designed for 377 of the sites repeated from the pilot library.

NGG-protospacers within these candidate enhancers were again scored using default parameters of FlashFry (McKenna and Shendure, 2018), and the third and fourth top scoring spacers were chosen to be used as an alternative gRNAs.

3.5.6 Choice of 948 exploratory candidate enhancers

Because the logistic regression classifier is biased toward the annotations that were used to select the initially targeted candidate enhancers (Figure 3.2A), we additionally used submodular subset selection to include DHSs optimized for a diversity of epigenomic features (Wei et al., 2015). We first removed from the full set of 29,833 DHSs (ENCSR000EKS) those 1,119 DHSs that were a part of the original screen. Note that we did not remove the 128 DHSs that had been selected again by the logistic regression model, because doing so would bias our remaining DHSs away from the same annotations. Then we calculated the Pearson correlation of overlapping epigenomic marks between the remaining DHSs. Lastly, we applied a facility location function (Mirchandani and Francis, 1990) to this similarity matrix and used a greedy submodular selection algorithm to identify 948 additional DHSs as exploratory candidate enhancers. The top two highest quality gRNAs (as scored by FlashFry) were included to target each candidate enhancer.

3.5.7 Note on choice of gRNA design for future screens of CRISPRi candidate enhancers

We used our set of enhancer-gene pairs to assess if there was a specific gRNA-target location within the candidate enhancer that increased CRISPRi efficacy. We correlated enhancer-gene pair effect size with each gRNA's absolute distance to center of either DHS-peak or overlapping p300 ChIP-seq peak. However, neither the absolute-distance-to-center-of-DHS-peak (Pearson's r: 0.02) nor the absolute-distance-to-center-of-overlapping-p300-peak correlated with effect size (Pearson's r: 0.07). Thus, we currently only recommend prioritizing gRNAs that fall within an open chromatin site based on quality and on-target efficiency as assessed tools like Flashfry (McKenna and Shendure, 2018).

3.5.8 gRNA-library cloning

The lentiviral CROP-seq gRNA-expression vector (Datlinger et al., 2017) was modified by Q5-Site Directed Mutagenesis (New England BioLabs, F:5-acagcatagcaagtttAAATAAGGCTAGTCCGTTATC-3 R:5-ttccagcatagctcttAAACAGAGACGTACAAAAAAG-3) to incorporate the previously described gRNA-(F+E)-combined backbone optimized for CRISPRi (Chen et al., 2013; Hill et al., 2018). Addgene #106280. Prepared vector was digested with BsmBI and alkaline phosphatase (FastDigest Esp3I and FastAP, Thermo Fisher Scientific), ”filler” sequence removed by gel extraction, and cleaned (Zymo Research DNA Clean and Concentrator-5) vector without ”filler” was used for all downstream cloning.

Spacer libraries were ordered as single stranded pools (CustomArray, 5-atcttgtggaaaggacgaaacaccGNNNNNN
3). 1 ng of each pool was amplified (F = 5-atcttGTGGAAAGGACGAAACA-3, R = 5-acttgctaTGCTGTTTCCAG
3, 64C Tm, Kapa Biosystems HiFi Hotstart ReadyMix (KHF), see Special note about gRNA-library cloning below, as we now recommended a different R primer = 5-CTGTTTCCAGCATAGCTCTTAAAC-3) and purified amplicons (Zymo Research DNA Clean and Concentrator-5) were cloned into CRISPRi-optimized CROP-seq vector prepared as described above (NEBuilder HiFi DNA Assembly Cloning Kit, NEB, 100 fmol purified vector: 200 fmol cleaned insert). 2 ul of each product was transformed into Stable Competent E. coli (NEB C3040H) in enough replicates to produce ¿ 20 transformant clones per gRNA in the library. Plasmid DNA was purified using ZymoPURE Maxiprep kits, following by DNA Clean and Concentrator cleaning (Zymo Research).

3.5.9 Special note about gRNA-library cloning

In Sanger sequence of the final gRNA plasmid libraries and in the 8-15 bp immediately downstream of the spacer (7 bp of the gRNA backbone transcript captured in all single-cell RNA-sequencing datasets), we identified that ˜80% of gRNAs harbored a small insertion or deletion (vast majority 1 bp deletions, Figure 3.14A) in between the spacer and the R primer 5-acttgctaTGCTGTTTCCAGC-3 used in the initial amplification of spacer-oligos. We inferred that this is due to slippage of the KHF polymerase as it copies the secondary structure of the first stem extension loop added as

part of the more stable sgOPTI backbone. In the scRNA-seq data, ~70% of gRNA carried a 1 bp deletion, ~8% carried a 2 bp deletion, and ~2% carried a 3 bp deletion (Figure 3.14A).

Fortunately, 1 bp deletions did not correlate with significant disruption of CRISPRi efficacy in the scRNA-seq data. (1 bp deletion % reduction) / (full length gRNA reduction) ratio was 1.01 (high confidence enhancer-gene pair) or 0.958 (TSS control). For 2 bp deletions, this ratio was also not extreme (0.959 (high confidence pair) or 0.806 (TSS control)). However, for 3 bp deletions (very rare), the ratio was 0.908 (high confidence pair) or 0.644 (TSS control). Overall correlation of all these deletion lengths to full length efficacy was very high (Figure 3.14B).

Thus, the vast majority (~90%) are either wild-type or harbor 1 bp deletions that create zero-to-little effect on CRISPRi efficacy. 8% of the remaining gRNA harbor 2 bp deletions that also largely do not affect CRISPRi efficacy. However, to avoid this problem in cloning future gRNA libraries into the sgOPTI-CROP-seq plasmid, we now recommend amplifying with a reverse primer that is flush with the spacer (5-CTGTTTCCAGCATAGCTCTTAAAC-3), potentially enabling a boost in repression efficacy.

3.5.10 Virus production and transduction

The Fred Hutchinson Co-operative Center for Excellence in Hematology Vector Production core produced all virus for the multiplexed enhancer-gene pair screening experiments. For the singleton CRISPRi recapitulation, virus was made in-house by co-transfecting (Lipofectamine 3000, ThermoFisher, L300015) HEK293Ts with the small pools of CRISPRi-optimized CROP-seq with the ViraPower Lentiviral Packaging Mix (ThermoFisher). After 3 days, supernatant was syringe filtered with a 0.45 uM filter (cellulose acetate, VWR) to prepare virus for transduction.

Cells were transduced (8 μg/mL polybrene) with varying titers and amounts of virus to achieve differing MOI. 400,000 and ~2.5 million original cells were transduced for the pilot and at-scale experiments, respectively. At 24 hours post-transduction, cells were spun and resuspended with virus- and polybrene- free media. At a total 48 hours post-transduction, 2 μg/mL puromycin was added to the culture, and changed to 1 μg/mL puromycin at the next passage for maintenance. A

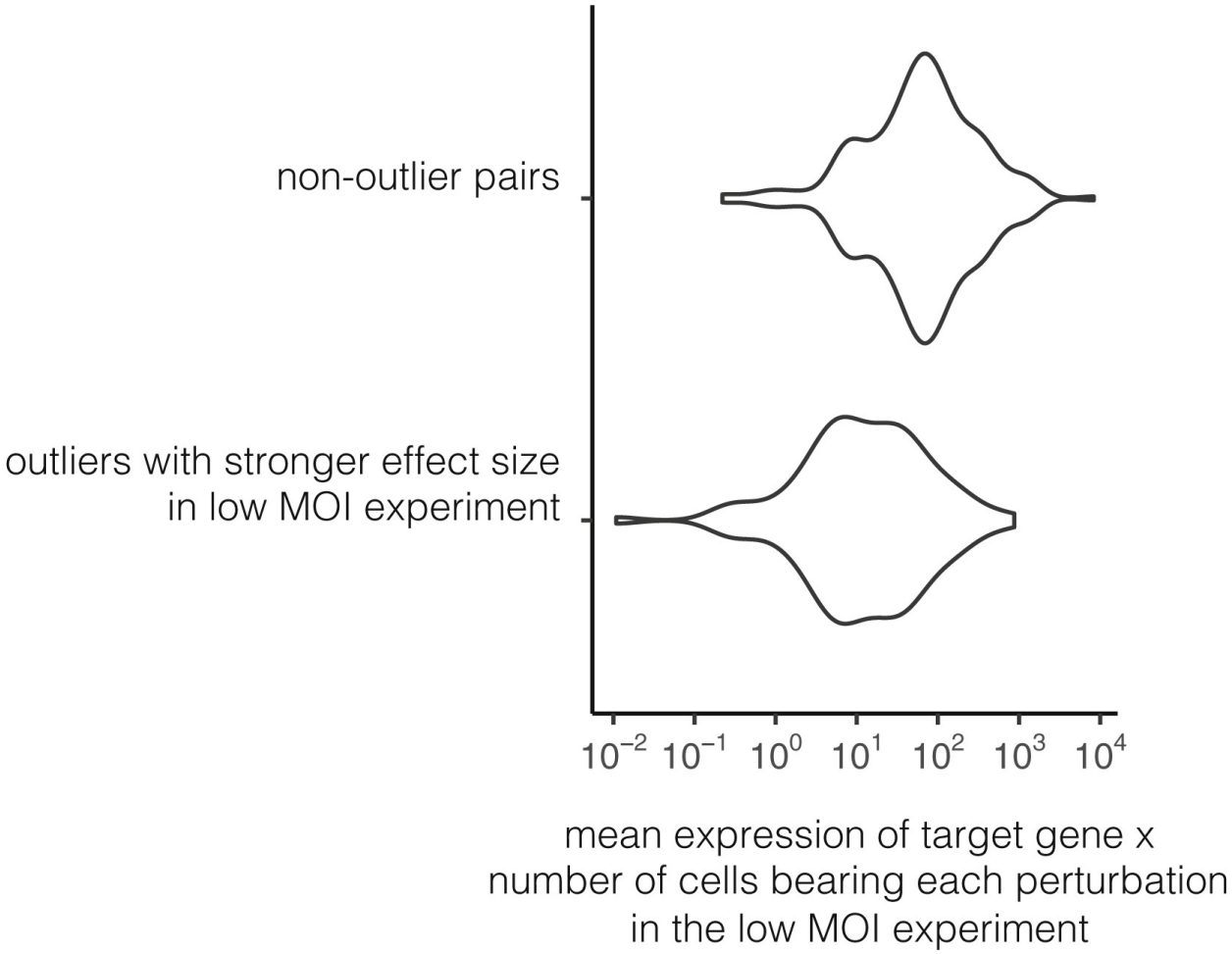


Figure 3.13: Outliers with Greater Effect Size in Low MOI Replicate Are Likely Due to Low Expression and Low Cell Count in Low MOI Replicate. The mean expression of the target gene in the low MOI 41,284 cell dataset as a function of the number of cells bearing each perturbation in that experiment.

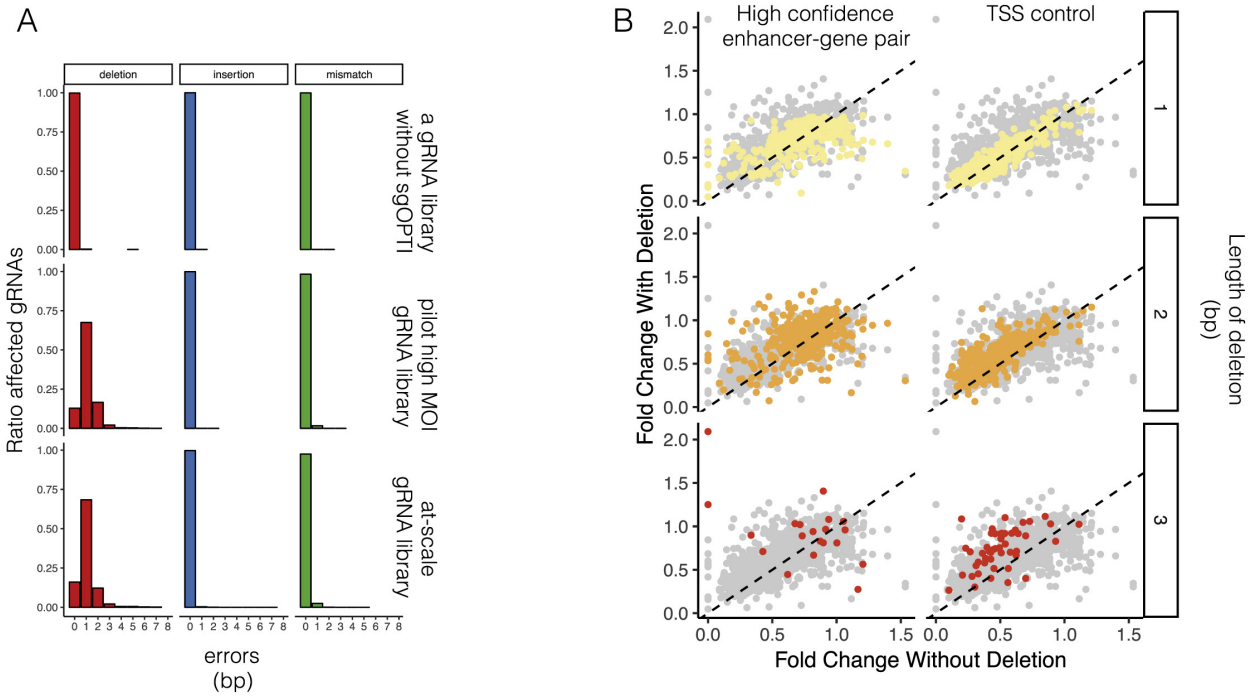


Figure 3.14: Supplementary Details, Related to STAR Methods. A and B) Quantification of errors in synthesis of the sgOPTI gRNA backbone across scRNA-seq datasets. (A) Deletion (red), insertion (blue), or mismatch (green) rate in the 8-15 bp downstream of the spacer in the gRNA backbone as captured by gRNA-transcript enrichment from scRNA-sequencing data. Data is shown for scRNA-seq datasets of a gRNA library that does not have sgOPTI added to the backbone (but was cloned, amplified and sequenced in a similar manner), the pilot high MOI gRNA library, and the at-scale gRNA library. (B) The impact of indels on effect sizes for paired-candidate enhancers (high confidence set) and TSS positive controls. The effect size of gRNAs with versus without perfect backbones, stratified by length of deletion. Gray points = a unique dot is plotted for the subgroups of each paired enhancer/TSS gRNA, divided by if they harbor 0, 1, 2, or 3 bp deletions. Colored points = set of gRNAs bearing the specified deletion length. Only points for which there are $\geq$ 50 cells in a given deletion-length group are plotted to ensure reasonable estimates of fold change.

total of 10 days post transduction, cells were collected for scRNA-seq or bulkRNA-seq.

3.5.11 Single cell transcriptome capture

˜4000-8000 cells were captured per lane of a 10X Chromium device using 10X V2 Single Cell 3' Solution reagents (10X Genomics, Inc). Six lanes were used for both the low and high MOI 1,119-pilot library experiments, and 32 lanes were used for the scaled experiment. All protocols were performed as per the Single Cell 3' Reagent Kits v2 User Guide (Rev B), except prior to the enzymatic shearing step, 100% of full length cDNA was taken for PCR enrichment of gRNA-sequences off the CRISPRi-optimized CROP-seq transcripts as described below. After RT, the 32 lanes of the scaled experiment were split into two batches (16 lanes each) for the remainder of the prep to enable easier handling.

3.5.12 gRNA-transcript enrichment PCR

A three-step hemi-nested PCR reaction was performed to enrich gRNA sequences from the 3' UTR of puromycin resistance gene transcripts produced by the CRISPRi-optimized CROP-seq integrant. PCR was monitored by qPCR to avoid overamplification, and each reaction was stopped immediately before it reached saturation.

In PCR 1 10-13 ng of full-length 10x scRNA-seq cDNA were amplified in each 50 μL KHF reaction (annealing temp 65C), spiked with SYBR Green (Invitrogen) for qPCR monitoring (10% of all unfragmented 10x cDNA) and used the following primers:

F: U6_OUTER 5- TTTCCCATGATTCCTTCATATTTGC -3 R: R1_PCR1 5- ACACTCTTTCCCTACACGAC
3

In PCR 2 sample replicates were pooled, cleaned with 1x Agencourt AMPure XP beads (Beckman Coulter), and 1/25th of the cleaned pooled product was amplified in a 50 μL KHF reaction spiked with SYBR Green and monitored as above (annealing temp 65C).

F: U6_INNER_with_P7_adapter 5-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGcTTGTGGAAAG
-3 R: R1-P5 5-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACG-3

In PCR 3 the PCR 2 replicate reactions were pooled and 1x AMPure cleaned. 1/25th of the cleaned pooled product was amplified in a 50 μL KHF reaction (spiked with SYBR Green and monitored as above, annealing temp 72C) and products cleaned once again via 1x Ampure.

F: 5-CAAGCAGAAGACGGCATACGAGATIIIIIIIIIIGTCTCGTGGGCTCGG-3 (standard NEXTERA P7 indexing primer where "I" represent index bases) R: R1-P5 again

3.5.13 Pilot library experiments sequencing

The final libraries were sequenced on a NextSeq 500 using four 75-cycle high-output kits (R1:26 I1:8, I2:0, R2:57) for each experiment (low and high MOI).

3.5.14 Scaled library experiments sequencing

The final library was sequenced by the Northwest Genomics Center on a NovaSeq 6000 using an S4 flow cell (R1:26, I1:8, I2:0, R2:91).

All libraries were sequenced to ˜20% sequencing saturation.

3.5.15 Digital gene expression quantification

Sequencing data from each sample was processed using the Cell Ranger software package as provided by 10x Genomics, Inc., to generate sparse matrices of UMI counts for each gene across all cells in the experiment.

Each lane of cells was processed independently using cellranger count, aggregating data from multiple sequencing runs. The pilot library experiments were each processed with cellranger 2.0.2; the at-scale library experiment was processed with cellranger 2.1.1.

3.5.16 Definition of genes well-expressed or "detectably expressed" in K562

Unless otherwise notes, genes were defined as well expressed or detectably expressed in K562 if they had at least one read in 0.525% of cells in their respective (pilot or at-scale screen) single cell RNA-seq datasets.

3.5.17 Assigning genotypes to cells

gRNAs were assigned to cells in the following method (Hill et al., 2018): Sequences corresponding to the gRNA-containing CRISPRi-optimized CROP-seq transcripts are extracted from the cellranger position sorted BAM file after running our custom indexed libraries through the cellranger pipeline to tag reads with corrected cell barcodes and UMIs. gRNA sequences are extracted and corrected to the library whitelist within an edit distance of two, and gRNA-cell pairs are tracked when a valid cell barcode and UMI are both assigned to the read. Likely chimeric reads are detected and removed to reduce noise in the assignments as previously described. We utilized thresholds to set minimum acceptable values for the total reads for a gRNA-cell pair and for the proportion of all CROP-seq transcript reads accounted for by each gRNA observed in a cell to distinguish noise from real assignments (Hill et al., 2018). Here, given the larger number of guides contained in each cell, we find that UMI counts provide a much cleaner distribution than read counts and have used UMI counts in all calculations. For the 1,119 pilot library experiments we used 0.01 read counts and 5 UMI in both our low and high MOI for each of these thresholds. For the scaled library experiment, we used 0.005 read counts and 5 UMI. Only cell barcodes that appear in the set of passing cells output by cellranger, which imposes an automated threshold on the total UMIs observed in cells, are carried forward in downstream analysis.

3.5.18 Differential expression tests

In our cis analyses, we tested each perturbing gRNA-group against genes within 1 Mb of the gRNA. These gRNA-gene pairs were identified by using bedtools to intersect the DHSs targeted by the gRNA library with 1 Mb windows in either direction of TSS annotations from GENCODE March 2017 v26lift37 (total of 2 Mb, centered around the TSS). In our trans analysis, all gRNA-groups were paired with all genes that were defined as expressed in K562. In both cis and trans analyses, NTCs were tested against any genes used to test perturbing-gRNAs.

For each gRNA-group we assigned a label of “1” to cells that contained a gRNA belonging to that group and a label of “0” to all other cells in the dataset. Monocle2 (Qiu et al., 2017a) was

used to perform a differential expression test, using the negbinomial.size family, over this categorical label to find differentially expressed genes between these two groups. Due to its support of complex model formulas, Monocle2 does not provide model coefficients as part of the differential expression results. We created a modified version of the differentialGeneTest function and associated helper functions that return both the intercept term and the coefficient of the group assignment to facilitate more robust prioritization and characterization of hits from our screen. The negative binomial family uses log as the link-function, so we can calculate the initial expression level as exp(intercept), and the fold change in expression between the two groups as exp(group_coefficient + intercept) / exp(intercept). We verified data from our power simulations that the appropriate effect sizes can be obtained with this method using the coefficients output by VGAM.

For the scaled experiment, as we collected a much larger number of lanes and observed the highest MOI, we regressed out the number of guide RNAs observed in a cell (as a proxy for the number of integrants), the percentage of total transcripts observed that are mitochondrial, and the prep batch (as following reverse transcription, the 32 lanes were prepared in two batches to make handling easier). In practice, we observe a modest boost in sensitivity when regressing out each of these factors in DE testing. This was done using the full model formula ~gRNA_group+guide_count+percent.mito+ and the reduced model ~guide_count+percent.mito+prep_batch in Monocle2.

3.5.19 Calling hits from differential expression test results

All differential expression test results were performed for all K562 expressed genes within 1 Mb of the target site as defined by GENCODE March 2017 v26lift37. NTCs were tested against all genes within 1 Mb of any target site.

Tests with two sources of potential false positives were excluded: 1. In the pilot experiment, we identified inflation of NTCs when testing them against genes highly impacted by perturbing-gRNA in our library (for example, NTCs associated with targets of our TSS and globin LCR controls). This was due to subtle yet detectable nonrandom associations of gRNA-groups with other gRNA-groups across cells, potentially due to slight bottlenecking at the transduction level

(400,000 cells transduced for 1,119 pilot library versus 2.5 million transduced for 5,779 scaled library). To exclude this source of inflation in the pilot dataset, we used Fisher's exact test to identify when an NTC was nonrandomly assorted with a perturbing-gRNA (adjusted P-value $<$ 0.01 and odds ratio $>$ 1). Then, any test of an NTC against a gene within 1 Mb of that gRNA's gRNA-group was excluded from further analytical steps.

2. We noted Monocle was susceptible to inflating P-values when a gene was highly expressed but only in few cells. Three of our 381 TSS controls fell into this category. To avoid this problem, we excluded outlier genes that were expressed in $<$ 20,000 cells in either the high-MOI 47,650-cell dataset and/or the scaled 207,324-cell dataset, and with log10(total UMIs / cells with a UMI) $>$ 0.2 greater than predicted by a spline fit generated via smooth.spline() with spar = 0.85 to limit overfitting (35 genes total).

Remaining tests were filtered to those that decreased expression of the target gene.

Then, an empirical P-value was defined for each gene-gRNA-group pair test as: [(the number of NTCs with a smaller P-value than that test's raw P-value) + 1] divided by [the total number of NTCs tests + 1].

These empirical P-values were Benjamini-Hochberg corrected, and those $<$ 0.1 were kept for 10% empirical FDR sets.

3.5.20 Use of 3.5% empirical FDR to initially select enhancer-gene pairs from the pilot study

We originally used an alternative method to call the original 145 enhancer-gene pairs from the original pilot dataset (a universal cutoff of the P-value at which the proportion of passing NTC-tests/total NTC-tests was 10% of the proportion of passing candidate enhancer tests/total candidate enhancer tests). However, upon further discussion and review of the eQTL literature, we revised our method to the one defined above. This original threshold corresponded to a 3.5% empirical FDR rate, as defined above.

3.5.21 Inclusive versus high confidence enhancer-gene pairs

The only requirement of enhancer-gene pairs in the inclusive set was that they passed a 10% empirical FDR in the scaled experiment. To be included in the high confidence set, enhancer-gene pairs either had to be replicated at a 10% empirical FDR in the pilot dataset, or (if a candidate enhancer was unique to the scaled experiment) both gRNAs had to be individually associated with ¿ 10% repression of the gene.

3.5.22 Analyses to evaluate reproducibility between gRNA

To evaluate reproducibility between gRNAs, we subset the 377 pairs (two sets of gRNA pairs targeting the same candidate enhancers in the scaled experiment) to pairs where both pairs negatively repressed at least one target gene (no significance requirement). 20 of the 377 did not meet this criteria. Then, we ranked all tested genes by average repression between the two gRNA pairs, and kept the top ranked gene for each pair. The repression levels of each type of gRNA pair on this top-ranked gene are plotted in Figure 3.5B, regardless of significance.

3.5.23 Intracellular abundance of gRNA and dCas9-KRAB transcript does not correlate with effect size

As both the dCas9-BFP-KRAB and the sgOPTI-CROP-seq construct transcripts are poly-A tagged, we are able to test if there is an association between the CRISPR components' UMI counts and transcript abundance of a targeted gene. For the 441 candidate enhancers in a high confidence pair, we subsetted to the cells that held a guide targeting each enhancer. Within this set of cells, we tested for a significant association between the expression of the target gene and the UMI count of the dCas9-BFP-KRAB or the guides (adjusting for total cell UMI count). Of the 470 enhancer-gene pairs, only 2 and 10 had any significant (adjusted P-value < 0.01; 7 and 27 for adjusted P-value < 0.05) association with dCas9-BFP-KRAB count or guide count respectively (0.4% and 2% or 1.5% and 5.7% of tests for each adjusted P-value threshold respectively). Based on this, we conclude there is not evidence for a substantial effect of dCas9-BFP-KRAB or guide counts on the observed

effect size for a given enhancer-gene pair.

3.5.24 Quantifying gRNA abundance

In the process of assigning gRNAs to cells, we had already quantified the number of reads and UMIs associated with gRNA-cell pairs. These counts were used as is for the above analysis.

3.5.25 Quantifying dCas9-BFP-KRAB in cells

We constructed a bowtie2 (Langmead and Salzberg, 2012) index for a reference including both the PuroR transcript from the sgOPTI-CROP-seq vector (extending from PuroR to the 3' LTR encoding the guide sequence as N's) and dCas9-BFP-KRAB (including the 3' LTR). Note that both gRNA and dCas9 transcripts were included in this analysis because several regions are identical within the 3' UTR of the transcripts encoded by these two constructs. We then took all the unmapped reads from the unbiased (cell) libraries and converted them back into fastq format adding the final cell ID and UMI from cellranger into the read name for use downstream. We mapped these reads to the reference above using bowtie2 using the command "bowtie2 -p 8 –n-ceil 20 –np 0 -x ¡reference¿ -U ¡fastq input¿ -S ¡bam output¿." We then took only reads that map uniquely to the dCas9 contig with mapq of 30 or greater and enumerated the number of UMIs and total reads seen for each cell / barcode pair dCas9.

In each case, we tested for associations between the gRNA/dCas9 counts and the abundance of each high confidence hit in our screen, only within cells that had a guide to the target. This was done using our modified version of differentialGeneTest as described above. Note that in this case we observed that size factors typically used to account for variation in total UMI counts across cells did not appear to sufficiently correct for the strong correlation between the counts of two transcripts (the gRNA transcript / dCas9 and the target) that results from variation in total UMI counts across cells. This initially resulted in residual associations that indicated increased gRNA transcript / Cas9 resulted in higher target expression. To account for this, we added an additional term to both the full and reduced model "~log10(total_umis)" and set all size factors to 1. This is

the model from which we report the above results.

3.5.26 Individual replication by CRISPRi singletons

To replicate a enhancer-gene pair's phenotype outside of the pooled mapping format, we prepared small pools of gRNAs re-targeting 15 high-confidence candidate enhancers or the TSSs of their respective paired-target genes. These enhancer-gene pairs were chosen from the following requirements: candidate enhancer tested in both the pilot and at-scale study (replicated between both); target gene in upper 50% of expression of all paired genes; target gene had no strong cancer associations or growth phenotypes. Additionally, we chose 6 candidate enhancers that were not paired with any target gene using the following requirements: tested in both the pilot and at-scale screen; empirical P-values for any cis gene ¿ 0.5 in both experiments; overlapping H3K27Ac ChIP-seq peak is in the top half of all the peaks that overlap the entire at-scale library (thus to be comparable with our paired enhancers); and within 1 Mb of a K562 expressed gene.

The two original gRNAs and two new gRNAs (making up the top 4 ranked on-target activity per candidate enhancer, filtering out those with high off-target scores using Flashfry (McKenna and Shendure, 2018); exception is candidate enhancer chr11.4680 where only 3 gRNAs passed these quality filters) were used for each respective pool, for a total of 4 gRNAs in the pool. The two original gRNAs were used for the TSS controls (plus two more alternative TSS gRNAs in the cases of NMU, GYPC, PTGER3, and PRKCB). These small gRNA pools were cloned into the CRISPRi-optimized-CROP-seq vector (as described above, except in the case of e-NMU targeting pool, which was cloned by ordering two reverse complement single stranded oligos and annealing them together into px459 (CRISPR-Reagent-Description_Rev20140509.pdf) (Cong et al., 2013). House-made lentiviral preps from these gRNA pools were transduced at low MOI into the K562-dCas9-BFP-KRAB line, and cultured for 10 days under puromycin selection before two technical replicates of total RNA were collected from each sample (RNeasy Mini Kit, QIAGEN).

Bulk RNA-seq libraries were prepared from each replicate via a TruSeq mRNA kit (400 ng input, Illumina, TruSeq RNA Sample Prep Kit v2 RS-122-2002 or TruSeq Stranded mRNA Library

Prep 20020595), and sequenced on a NextSeq 500 (total two 150-cycle kits cycling 80/80/6 in mid output mode for e-NMU, e-PRKCB, e-GYPC, e-PTGER3; total two 75-cycle kits cycling 40/40/8 in high output for all others; aiming for 10-20 million reads/sample). Gene-level quantifications and differential expression tests were performed via kallisto (Bray et al., 2016) and sleuth (Pimentel et al., 2017). Repression percentages were calculated from the kallisto transcript per million output table (normalized by size factors): (mean between the two replicates / mean between all-non targeting samples). To note, targeting the TSS of CITED2 did not seem to successfully repress CITED2's expression, though this is potentially due to inaccuracy of 1 of 2 technical replicates for this sample. The 3 that matched direction and magnitude of effect but were not significant in a test of differential expression potentially were not detectable due to lack of power, as we sequenced only two RNA replicates per sample.

To note: we additionally generated singleton datasets for chr6:34191315-34191338 (paired with HMGA1 in the pilot screen), but did not include this in analysis as it did not reproduce between the pilot and at-scale screen, and thus was not part of our high confidence enhancer-gene pair set.

3.5.27 Validation by sequence deletions

To generate monoclonal sequence lines of three candidate enhancers, we designed protospacers to flank the DHSs targeted in e-NMU, e-GLUL, and e-CITED2. Spacers were order as single stranded oligos (IDT) and then amplified (KHF, 5-GTGGAAAGGACGAAACACCg-3, 5-gctaTTTCtagctctaaaac-3, 55C tm, 15 s extension; followed by clean-up via Zymo Research DNA Clean and Concentrator) to be made double stranded for Gibson Assembly cloning (50 ng prepared vector: 0.66 ng prepared insert) into the Cas9- and gRNA- expression vector px459 (Ran et al., 2013), expressing both the gRNA and a cassette of Cas9-2A-puromycin resistance; NEBuilder HiFi DNA Assembly Cloning Kit). Some e-NMU targeting oligos were cloned by annealing two complementary oligos together followed by ligation into px459, in the method of CRISPR-Reagent-Description_Rev20140509 (Cong et al., 2013).

We transiently transfected the small px459 pools into the K562+dCas9-KRAB cell line using the Neon nucleofection system (500,000 cells per library, 10 μL tips, 500 ng of plasmid, pulse voltage 1450–pulse width 10–pulse number 3; ThermoFisher). Beginning 24 hours after transfection, cells were selected with 1 ug/mL puromycin for 48-72 hours, then single-cell sorted into 96 well plates using a FACSAria II (Becton Dickinson). To finally achieve clones that harbored fully homozygous deletions of e-NMU, this process was repeated on an initial set of heterozygous clones using a second round of flanking gRNAs.

After 3-4 weeks of growth, gDNA was extracted by concentrating cells into 20 μL of media, and adding 40 μL of house-made Quick Extract buffer (EB + 4 mg/mL proK + 0.45% Tween20), followed by 65C for 6 minutes and 98C for 2 minutes. 1 μL of this gDNA extract was used for genotyping PCRs (Kapa2G Robust PCR kit, 35 cycles 60C-HS-3 minute extensions).

Two rounds of genotyping PCRs were performed. First, clones were screened with primers flanking the deletion to identify clones that harbored a deletion on at least one allele. Second, to confirm homozygosity, primers internal to the deleted region were used to identify candidates that still harbored wild-type alleles (Figure 3.8A). Clones that harbored full deletions with no remaining wild-type alleles were submitted to bulkRNA-sequencing (Figures 3.6E–G). Two technical replicates of RNA were extracted from each monoclonal line (RNeasy Mini Kit, QIAGEN), bulkRNA-seq libraries prepared via a TruSeq mRNA kit (400 ng input, Illumina, TruSeq Stranded mRNA Library Prep 20020595), and sequenced on a NextSeq 500 (one 75-cycle kits cycling 40/40/8 in high output for monoclonal samples; aiming for 10-20 million reads/sample). Gene-level quantifications were performed as for the CRISPRi singletons, and reduction percentages calculated from kallisto transcript per million output table (normalized by size factors): (mean of all replicates per candidate enhancer) / (mean between all-non targeting samples).

3.5.28 Phenotyping e-NMU perturbations by flowFISH

Cells harboring e-NMU CRISPRi perturbations were generated as in Individual replication by CRISPRi singletons. A heterogeneous population of cells harboring full e-NMU deletions was

generated as in Validation by sequence deletions (though without single-cell clone sorting). A heterogeneous population of cells harboring scanning deletions across e-NMU was generated by cloning and transfecting 19 gRNAs targeted every ˜100 bp across the e-NMU locus as described above in Validation: sequence deletions.

Fluorophore labeled complementary probes to NMU transcript were designed on and ordered from `https://www.molecularinstruments.com/`. The 'non-targeting' probes were scrambled versions of the original NMU-targeting probes (to preserve sequence features such as GC content). RNA flowFISH was performed according to Molecular Instruments' in situ HCR v3.0 protocol (Choi et al., 2018), which we have described again here: Cells were by resuspending in 4% formaldehyde to reach 10ˆ6 cells/mL, and fixing for 1 hour. Formaldehyde was then removed, cells were washed four times in PBST (1x PBS + 0.1% Tween 20), and then resuspended in 70% ethanol. For labeling, cells were first washed twice with PBST, and then pre-hybridized by incubating at 37C for 30 minutes in 30% probe hybridization buffer (30% formamide, 4x sodium chloride sodium citrate (SSC), 9 mM citric acid, 0.1% Tween 20, 50 ug/mL heparin, 1x Denhardt's solution, and 10% low MW dextran sulfate). Cells were then incubated overnight at 37°C in a final 4 nM probe solution (prepared by adding 2 pmol each probe (a mix of 1 uL of 2 uM stock per each probe) + 100 uL of 30% probe hybridization buffer). Cells were then repeatedly resuspended in 30% probe wash buffer and incubated for 10 minutes at 37°C, for a total of four washes. Cells were then resuspended in 5x SSCT (5x SSC + 0.1% Tween 20) and incubated at room temperature for 5 minutes before amplification.

For amplification, cells are resuspended in amplification buffer (5x SSC + 0.1% Tween 20 + 10% low MW dextran sulfate) and pre-amplified by incubating for 30 minutes at room temperature. 15 pmol of each fluorescently labeled hairpin was snap-cooled by heating 5 uL of 3 μM stock in hairpin storage buffer (Molecular Instruments) to 95C and then cooling for 30 minutes to room temperature in a dark drawer. Snap-cooled hairpins were then mixed with amplification buffer, added to the sample for a final concentration of 60 nM, and then incubated overnight ($>$ 12 hours) at room temperature in the dark. Cells were washed six times by resuspension in 5x SSCT, resuspended in 500 μl 2x SSC, and incubated for 30 minutes at room temperature with 0.5 μL

Vybrant Dye Cycle Orange (DNA stain).

For sorting, the cells are first gated based on size and granularity using forward versus side scatter to discriminate between debris and cells. Cells in G0/G1 stage are then selected using DNA dye (Vybrant Dye Cycle Orange).Cells are then sorted into low, medium, or high bins of NMU expression using AF647 (Becton Dickinson; ˜500,000 cells for the full deletion low NMU bin, ˜1,000,000 cells for all other bins).

To reverse cross-link the sorted samples, cell pellets were resuspended in 500 μl of elution buffer (4 mL H2O + 500 μl 10% SDS + 500 ul NaHCO3 1M) + 30 ul of NaCl (5M) and incubated overnight at 65C. 8 ul of RNase (10 mg/mL) was added to each sample, mixed by inversion, and incubated at 37C for 2 hours. 4 μl of Proteinase K (20 mg/mL) was added, mixed by inversion, and incubated for 2 hours at 55C. gDNA was extracted by phenol chloroform, ethanol precipitated, and resuspended in QIAGEN elution buffer.

PCR to identify e-NMU genotype enrichments in each of the NMU expression bins (Figures 3.8B and C) was performed using Kapa2G Robust (e-NMU outer PCR: F primer 5'TCCAACC-CCTCAACTTGTT3' Reverse primer 5'TGCCTTCTCTGCCTTTCATT3'; anneal 60C, extension time 1:50) on 10 ng of gDNA. PCRs were spiked with SybrGreen, and monitored on a qPCR to allow removal before overamplification to prevent excessive PCR biases. 1 uL of each PCR reaction was run on a 6% TBE polyacrylamide gel (Invitrogen) for 35 minutes at 180 V and stained with Sybr Gold for visualization. Replicate PCRs are represented by different lanes in Figures 3.8B and C.

3.5.29 Aggregate analysis of enhancer-gene pairs

The high confidence enhancer-gene pairs were used for all the following analyses unless otherwise noted. Details of empirical FDR and the significance thresholds used to call enhancer-gene pairs can be found above in Calling hits from differential expression test results. Singleton re-testing and validations of enhancer-gene pairs used to functionally test if the data met the assumptions of these statistical methods can be found above in Replication of enhancer-gene pairs as singletons.

3.5.30 Distance between perturbation and target gene

Distance was calculated between the GENCODE March 2017 v26lift37 annotated TSS of the perturbed gene and the middle of the originally targeted open chromatin region (if targeting a candidate enhancer, ENCFF001UWQ) or the GENCODE-annotated TSS of the originally targeted transcript (if targeting a TSS). To note, in Figure 3.10A and to calculate the median distance, we have only used enhancers that are upstream of the target gene, as the length of the gene body would confound distance-to-TSS measurements for downstream enhancer-gene pairs.

3.5.31 Expression distributions

Average expression of each transcript was defined as mean UMI counts per cell in the 47,650 or 207,324 cell scaled dataset. K562 expressed genes were defined as at least one read in 0.525% of cells in the same dataset.

3.5.32 ChIP-seq strength quintile analysis and logistic regression classifier

All candidate enhancers targeted in each library were bedtools-intersected with 170 ChIP-seq of histone-associated marks (ENCODE Project Consortium, 2012)), broken into quintiles of the 7th "signalValue" column (peak strength, usually representing overall average enrichment in the region), and the rates of enhancer-gene pairs identified in each quintile were used. In addition to average phyloP conservation score per candidate enhancer, these were used to fit both independent and multivariate logistic regression classifiers using the glm() function with binomial family. We calculated fold changes for how likely a candidate enhancers was paired by: 1 + (((odds ratio - 1) $*$ highest quintile ChIP-seq value) - ((odds ratio - 1) $*$ lowest quintile ChIP-seq value)).

3.5.33 Motif enrichment in enhancers and promoters

Using the AME tool (Analysis of Motif Enrichment) from the MEME suite (McLeay and Bailey, 2010), enhancer analysis: we compared motifs enriched in the 600 candidate enhancers in the

inclusive set of 664 pairs as compared to all 5,779 in the at-scale library; promoter analysis: compared motifs enriched the 1 Kb upstream of the TSS (˜promoter) of the 479 genes in the inclusive 664 pairs as compared to the ˜promoters of all K562 expressed genes within 1 Mb of a tested candidate enhancer. Parameters were set to default, and Hocomoco Human v11 (core) (Kulakovskiy et al., 2013) was used as the motif library.

3.5.34 Motifs of TF couples across paired promoters and enhancer

To test if pairs of transcription factor (TF) motifs were enriched for co-presence across paired promoters and enhancers, we first identified 179 TFs that were expressed in K562s and had high quality motifs in Hocomoco. Using the FIMO tool (Find Individual Motif Occurrences) from the MEME suite, we annotated all 600 candidate enhancers and the promoters of all 479 genes (1 Kb upstream of the TSS) in the inclusive set of 664 pairs. Motifs in the bottom quartile of how often seen in a promoter were excluded for lack of power. Then, we looped through all possible pairs of 179 TFs in the enhancer (TFe) x 179 TFs in the promoter (TFp), and for each TFe x TFp pair, performed a Fisher's Exact test on contingency tables designed as follows:

For the promoters of 479 paired genes: TFp in promoter or TFp not in promoter versus Promoter paired with an enhancer that contains TFe or Promoter not paired with an enhancer that contains TFe

For the 600 paired enhancers: TFe in enhancer or TFe not in enhancer versus Enhancer paired with an enhancer that contains TFp or Enhancer not paired with an enhancer that contains TFp The six TFe x TFp co-enriched couples that had a Benjamini Hochberg corrected P-value < 0.1 for both the 479 paired promoter analysis and the 600 paired enhancers analysis were described in the main text and supplementary data files.

3.5.35 ChIP-seq of TF couples across paired promoters and enhancer

Bedtools was used to mark when a paired enhancer or promoter in the 664 inclusive dataset overlapped a ChIP-seq peak from ENCODE generated K562 datasets were used. ChIP-seq datasets

that that were in the bottom quartile of how-often-overlapping with a paired enhancer or promoter were excluded for power (leaving 168 TFe and 166 TFp). Analysis was then performed the same as in the TFe x TFp motif analysis (Fisher's Exact Test, adjusted P-value < 0.1, pair required to be enriched when looping through both enhancers and then through promoters, TFe and TFp required to be different).

3.5.36 Functional annotation enrichment

We used the Piano package (Väremo et al., 2013) to perform functional annotation enrichment from the 'all pathways' Gene Ontology (`http://download.baderlab.org/EM_Genesets/June_20_2014/Human/June_20_2014_versions.txt`). The 10,560 K562-expressed genes within 1 Mb of a perturbing-gRNA were used as our background dataset, and randomly sampled from genes with expression greater than one standard deviation below the mean of our 353 targeted genes was used as the comparison set of "expression matched controls" (Figure 3.10C).

3.5.37 Hi-C analysis

We used the in situ Hi-C dataset for K562 cells from Rao et al. (2014), using the MAPQ 0 threshold and KR normalization, at 5 Kb resolution. We first created shuffled control loci pairs by starting with the set of enhancer-gene TSS pairs, and randomly shuffling the oriented distances between enhancer-TSS pairs, keeping either the enhancers or the TSSs intact. The rare cases where shuffling resulted in an invalid chromosomal coordinate were excluded. For each set of loci pairs, we identified the TADs (as defined inRao et al. (2014) using Arrowhead) encompassing each loci pair. For overlapping domains, we used the farthest domain boundary on each side of the loci pair. We omitted loci pairs that were not encompassed by any TADs from further analysis. We then extracted the normalized Hi-C counts for each loci pair, along with those for all other bins representing interactions at the same genomic distance within the same TAD, and calculated its fractional rank (scaled from 0 to 1, with 1 representing the highest interaction frequency). Finally, the distributions of fractional ranks were plotted and compared. In addition to comparing interac-

tions within TADs, we also compared loci pairs to other bins within 200 Kb or 1 Mb of each loci pair.

3.5.38 Analyses for multiplexability of CRISPRi within cells - low versus high MOI comparisons

In order to confirm the efficacy of repression in our high MOI experiments (pilot library MOI = ~15 and at-scale library MOI = ~28), we sought to compare the degree of repression observed in each of these experiments to that observed in our low MOI control experiment (pilot library MOI = ~1). We took all gene-target site differential expression tests passing a 10% empirical FDR in any one of the three experiments (as evaluated independently in each screen). We used this set rather than our final hit list to ensure that we were not biasing our comparison by excluding tests that would be independently called by any one screen but not the others, although we note that the results of the same set of analyses using our final set of hits are very similar.

For each of these tests, we calculated the observed fold changes of repression (where 1 is no change and 0 is complete loss of expression) for each screen and then calculated the following ratios: (pilot high MOI fold change) / (pilot low MOI fold change) and (at-scale fold change) / (pilot low MOI fold change), using a pseudocount of 0.01. As we found it potentially confusing that a higher value of these ratios represents worse efficacy of repression in the high MOI experiments, we considered making these ratios of percent repression (1 - fold change). However, as this value could be negative in some cases (where the fold change was greater than one in one of the screens), this was not compatible with display on a log scale. Therefore, in all plots showing such ratios, we are actually showing the inverse of the fold change ratios described above, which should approximately represent the ratios of percent repression without producing any negative values. Thus, in our plots and reported summary statistics, values less than one represent cases where more repression was observed in the low MOI control.

Despite the distributions of the ratios described above being centered at one, which indicates largely equivalent repression in high and low MOI experiments, there was a left tail, representing a smaller number of tests with reduced estimated efficacy in the high MOI experiments. We reasoned

that this could be an artifact of these genes being more lowly expressed and/or being represented by fewer cells given the sparse sampling of the pilot library in the low MOI experiment. In either case we might tend to underestimate the amount of transcript remaining after repression or at the very least the estimates would be substantially noisier, resulting in an artifactual tail. To confirm the lower expression levels genes in the observed tail, we took all tests falling in the first quartile of each distribution and compared the expression of these genes (average expression for the pilot low MOI experiments in the group of cells without the relevant gRNA; calculated by exponentiating the intercept from the differential expression test, which in the pilot high differential test is the estimated expression in UMI counts for the group of cells without the relevant gRNA). We further scaled these values by the total number of cells observed for each gRNA group in the pilot low MOI experiment to examine the combined effect of representation and expression level, which both contribute to what we expect is simply less robust estimation of fold change. We note that this scaling does not appreciably impact the overall distributions in this case.

3.5.39 Power simulations

In order to predict the impact of multiplexing on the power of enhancer-gene pair screens, we developed a simulation framework. First, using single-cell RNA-seq data collected from the pilot 47,650 K562 cells, we estimated a dispersion function that relates the mean expression of a gene to its dispersion estimate (one of the two parameters required for the negative binomial distribution) calling the Monocle2 functions estimateSizeFactors and estimateDispersions. This function is typically used in differential expression testing to shrink dispersion estimates, but here we use it to estimate dispersion values for simulated transcripts. This dispersion function is then extracted from the CellDataset object output by Monocle2 and used as input to our simulations.

Next, we chose relevant ranges for each of the parameters varied in our simulation: the MOI, total cell count, effect size (fraction repressed by CRISPRi), and mean expression level of the gene being tested. By examining the range of expression values observed in our data, we chose to simulate expression data for genes having mean expression values (size parameter of the nega-

tive binomial distribution) of 0.01, 0.1, 0.32, 1.0, 3.16, and 10.0 UMIs (0.10, 0.32 and 1.00 used respectively as low, medium, and high in Figure 3.1B) to provide a range of representative values.

We simulated MOIs at several values from 0.3 to 50, a range which includes the MOIs estimated from our own enhancer-gene pair screens. For each MOI, we calculate the expected number of cells containing a given guide by assuming a Poisson distribution of lentiviral delivery, zero-truncating the distribution to account for drug selection for cells that contain a guide transcript, and rescaling the probability distribution of guide counts accordingly. Perfect library uniformity was assumed to obtain the expected number of cells containing a given guide and the number of cells that do not contain that guide. Effect sizes of CRISPRi repression were chosen using estimates from the literature and were simulated at several values between 10% to 90% percent repression of the average expression level of the target transcript (size parameter input to the negative binomial distribution).

Finally, we simulated several values of total cells included in the experiment ranging from 35,000 to 300,000 cells (45,000 cells shown in Figure 3.1B). Expression data from transcripts corresponding to 100 samplings per set of parameters were generated for the populations of cells containing the gRNA and not containing the gRNA respectively. Our expression data simulation assumed a negative binomial distribution with the appropriate size parameter for the cells with and without the gRNA, and a dispersion value estimated using the dispersion function described above given the starting mean expression level being simulated. For each set of parameters, the simulated transcripts were subjected to a differential expression test performed between cells with and without the gRNA assigned using our modified version of the Monocle2 function differentialGeneTest as described above (see Differential Expression Tests). P-values were obtained and corrected assuming an average number of 20 tests per group in the library to approximate the number of genes contained within 1 Mb on either side of each gRNA-group and the impact of multiple testing. The rate of tests falling below a adjusted P-value of 0.05 were tabulated at each set of parameters to make power curves.

3.5.40 Quantify errors in gRNA backbone as described in Method Details: "Special note about gRNA-library cloning"

To quantify the rate of mismatches and indel lengths in the gRNA backbones for each library, we extracted the backbone portion of the gRNA transcript for each read in our gRNA transcript enrichment libraries and aligned it to the expected reference, (gtttAagagctaTGCTGGAAACAGCAtagcaagttTaaat), using semi-global version of the Needleman-Wunsch algorithm implemented by RecNW (Yahi et al., 2018). Mismatch and indel counts were made within the hairpin portion of the backbone (we initially screened backbone bases 8 to 31 downstream of the spacer), to restrict to bases that would be the most likely to have some if any functional impact. However, it should be noted that the overwhelming majority of all indels were small deletions observed in bases 8 to 14 or so; thus, rates provided in Figure 3.14A are limited to these 7 bp. For the pilot-gRNA libraries, where we had a shorter cDNA read length that does not cover the entire hairpin, so we simply quantified mismatches and indels in the 8 to 14 bp window (which again contained the overwhelming majority of all indels in our at-scale gRNA library). For each target-UMI pair in each cell, we averaged the observed mismatch and indel counts/lengths to get a consensus over all reads with a given UMI. We then averaged the statistics derived from UMIs for each target-cell assignment to get a final set of statistics for each. Each average was rounded to the nearest integer for plotting. This allowed us to quantify rates across screens and also examine how any changes in effect sizes correlated with effect sizes.

3.5.41 tSNE clustering of each dataset to check for biological distortions

We tested for enrichment of gRNAs in specific tSNE-based clusters of the at-scale single cell transcriptome dataset, to identify any perturbed targets that resulted in stronger changes to global expression, presumably mediated through trans effects of the target gene. For the at-scale dataset, we subsetted to genes that were expressed in at least 0.5% of cells and 50,000 cells were randomly sampled. We then processed the dataset using Seurat (Butler et al., 2018). We removed cells with greater than 10% mitochondrial transcripts, ran NormalizeData, and found the top 5,000 variable

genes using FindVariableGenes. Using these top 5,000 variable genes as input we then ran ScaleData, regressing out the percent of each cell's transcriptome accounted for by mitochondrial genes. We then computed 100 PCs using RunPCA (weighting PCs by variance explained), which were used as input to both the FI-tSNE method using RunTSNE and Louvain clustering at a resolution of 0.5 using FindClusters. Fisher's Exact tests were performed to test for a perturbed target's enrichment in each cluster. 8 TSS controls and 6 candidate enhancers were enriched within specific clusters (odds ratio ¿ 5, adjusted P-value < 0.01). However, even in these cases, only 10% of cells in which the target is perturbed actually fall into the cluster in which they are found to be enriched. Thus, this is not expected to compromise the screen, as in order to be a chronic source of false positives, the gRNAs targeting these global-change genes would have to be non-randomly associated with other gRNAs in the library.

3.6 Data availability

The accession number for the sequencing data (single cell RNA-seq and bulkRNA-seq) and processed data files is GEO: GSE120861 (metadata file), and GSM3417251–GSM3417303 (actual datasets).

Differential expression code and description of relevant data files provided in GEO are provided at `https://github.com/shendurelab/tafka-crisprQTL`. In addition, several supplemental tables are provided as part of the supplementary material for Gasperini et al. (2019).

3.7 Project acknowledgments

We thank the entire Shendure, Trapnell, and Ahituv labs, in particular D. Cusanovich, V. Agarwal, J. Tome, G. Findlay, S. Domcke, S. Regalado, J. Klein, L. Starita, D. Aghamirzaie, A. McKenna, X. Qui, F. Inoue, and O. Elor. We additionally thank J. Ousey, K. Han, M. Bassik, and M. Kircher. This work was supported by awards from the NIH (UM1HG009408 to N.A. and J. Shendure, DP1HG007811 to J. Shendure, and U24HG009446 to W.S.N.), the W.M. Keck Foundation (to C.T. and J. Shendure), and training awards from the National Science Foundation and NIH (Grad-

uate Research Fellowship to A.J.H. and M.G. and 5T32HG000035 to M.G.). J. Shendure is an Investigator of the Howard Hughes Medical Institute.

Chapter 4: A SINGLE-CELL ATLAS OF IN VIVO MAMMALIAN CHROMATIN ACCESSIBILITY

Chapter 4 is adapted with minimal modification from:
Cusanovich, D.A.*, Hill, A.J.*, Aghamirzaie, D., Daza, R.M., Pliner, H.A. Berletch, J.B., Filippova, G. N., Huang, X., Christiansen, L., DeWitt, W.S., Lee, C., Regalado, S.G., Read, D.F. Steemers, F.J., Disteche, C.M., Trapnell, C., and Shendure, J. (2018a) A Single-Cell Atlas of *In Vivo* Mammalian Chromatin Accessibility. Cell. *174*, 1309–1324.

4.1 ABSTRACT

We applied a combinatorial indexing assay, sci-ATAC-seq, to profile genome-wide chromatin accessibility in ~100,000 single cells from 13 adult mouse tissues. We identify 85 distinct patterns of chromatin accessibility, most of which can be assigned to cell types, and ~400,000 differentially accessible elements. We use these data to link regulatory elements to their target genes, to define the transcription factor grammar specifying each cell type, and to discover in vivo correlates of heterogeneity in accessibility within cell types. We develop a technique for mapping single cell gene expression data to single-cell chromatin accessibility data, facilitating the comparison of atlases. By intersecting mouse chromatin accessibility with human genome-wide association summary statistics, we identify cell-type-specific enrichments of the heritability signal for hundreds of complex traits. These data define the in vivo landscape of the regulatory genome for common mammalian cell types at single-cell resolution.

4.2 INTRODUCTION

Efforts to produce a human cell atlas are in their infancy, challenged in part by the fact that an adult human (70 kg) consists of a staggering ~37 trillion cells (Bianconi et al., 2013). Furthermore, cell types vary in abundance by several orders of magnitude and occupy a range of cell states over the course of development. The house mouse, Mus musculus, is the foremost model organism for

biomedical research. By extrapolation, an adult mouse (20 g) consists of a mere ˜10 billion cells. Mouse strains are isogenic and can be genetically manipulated. Coupled with the fact that cell types may be best understood through the lens of evolution (Arendt et al., 2016), we are motivated to pursue a cell atlas of mouse.

Technologies for the molecular profiling of single cells are diversifying, but recent organism-scale cell atlases (Cao et al., 2017; Fincher et al., 2018; Han et al., 2018; Karaiskos et al., 2017; Plass et al., 2018; The Tabula Muris Consortium et al., 2017) all use single-cell RNA sequencing (scRNA-seq). Although powerful for disentangling cell-type heterogeneity in complex tissues, scRNA-seq fails to capture the chromatin regulatory landscape that governs transcription in each cell type.

Chromatin accessibility is a generic marker of regulatory DNA classically measured by DNase I hypersensitivity, now read out by sequencing (DNase-seq) (Hesselberth et al., 2009). ATAC-seq (Buenrostro et al., 2013) is an alternative to DNase-seq, measuring chromatin accessibility to insertions by the Tn5 transposon (Adey et al., 2010). There have been several large-scale efforts to map genome-wide chromatin accessibility in cell lines and tissues (Roadmap Epigenomics Consortium et al., 2015; Thurman et al., 2012). However, cell lines are altered by in vitro culturing and tissues confounded by cell-type heterogeneity. Although in vivo cell types can be flow sorted and studied, this is labor intensive and requires a priori knowledge of markers.

We recently adapted combinatorial indexing (Amini et al., 2014) to single cells (Cusanovich et al., 2015). With single-cell combinatorial indexing (sci-), nucleic acids from each of many cells are uniquely tagged through several rounds of "split-pool" barcoding. To date, we and colleagues have developed sci- protocols for chromatin accessibility (sci-ATAC-seq) (Cusanovich et al., 2015, 2017), transcription (Cao et al., 2017), genome conformation (Ramani et al., 2017), DNA sequence (Vitak et al., 2017), and DNA methylation (Mulqueen et al., 2017).

Here, we set out to generate a single-cell atlas of in vivo mammalian chromatin accessibility. We applied sci-ATAC-seq to measure chromatin accessibility in ˜100,000 single cells derived from 17 samples representing 13 tissues of adult mice. From these data, we identify diverse cell types, define candidate tissue-specific enhancers, and model the transcription factor (TF) regula-

tory grammar that specifies each cell type. We also use these data to link distal regulatory elements to their target genes, characterize in vivo heterogeneity in chromatin accessibility within cell types, and identify cell types of principal relevance for common human diseases and traits.

4.3 RESULTS

We isolated nuclei from 13 distinct tissues of 8-week-old male C57BL/6J mice (Figure 4.1A). For four tissues, we collected a replicate sample from a second mouse. Nuclei were processed through an optimized sci-ATAC-seq protocol in batches, and all libraries were sequenced as a single pool (Figure 4.1B). For quality control (QC), we examined whether our tissue-level ATAC-seq data (prior to splitting into single cells) were reproducible between replicates and correlated with DNase-seq data from ENCODE (Yue et al., 2014). We first identified peaks of accessibility in each tissue (Figure 4.2A) and then evaluated quantitative measures of accessibility between different samples across the union of all peaks. Although the proportion of reads overlapping peaks was lower for our data (median 0.36 for sci-ATAC-seq versus 0.44 for ENCODE DNase-seq; 4.2B), the four tissues that we profiled in replicate were well correlated (Spearman's rho from 0.89 to 0.94 for ATAC-seq replicates versus 0.66 to 0.92 for DNase-seq replicates, Figures 4.2C–I). In hierarchical clustering, most samples from the same tissue tended to cluster together regardless of the assay used to generate the data (Figure 4.2J).

With ˜3.9 billion read pairs, we identified 104,039 cells, with ˜12% of these being likely doublets or "collisions" (Cusanovich et al., 2015). The total number of cells profiled per tissue (after further filtering detailed below) ranged from 2,278 for cerebellum to 9,996 for lung (Figure 4.1C). The minimum unique read depth acceptable per cell was determined for each tissue and ranged from 656 for one bone marrow replicate to 1,734 for thymus (Figure 4.1D, left). The number of unique reads per cell varied, ranging from a median of 8,743 for cerebellum to 23,456 for the prefrontal cortex (Figure 4.2K). We estimate that the current sequencing depth accounts for between 38% (testes) and 90% (one lung replicate) of the unique reads present in each library (Figure 4.2L).

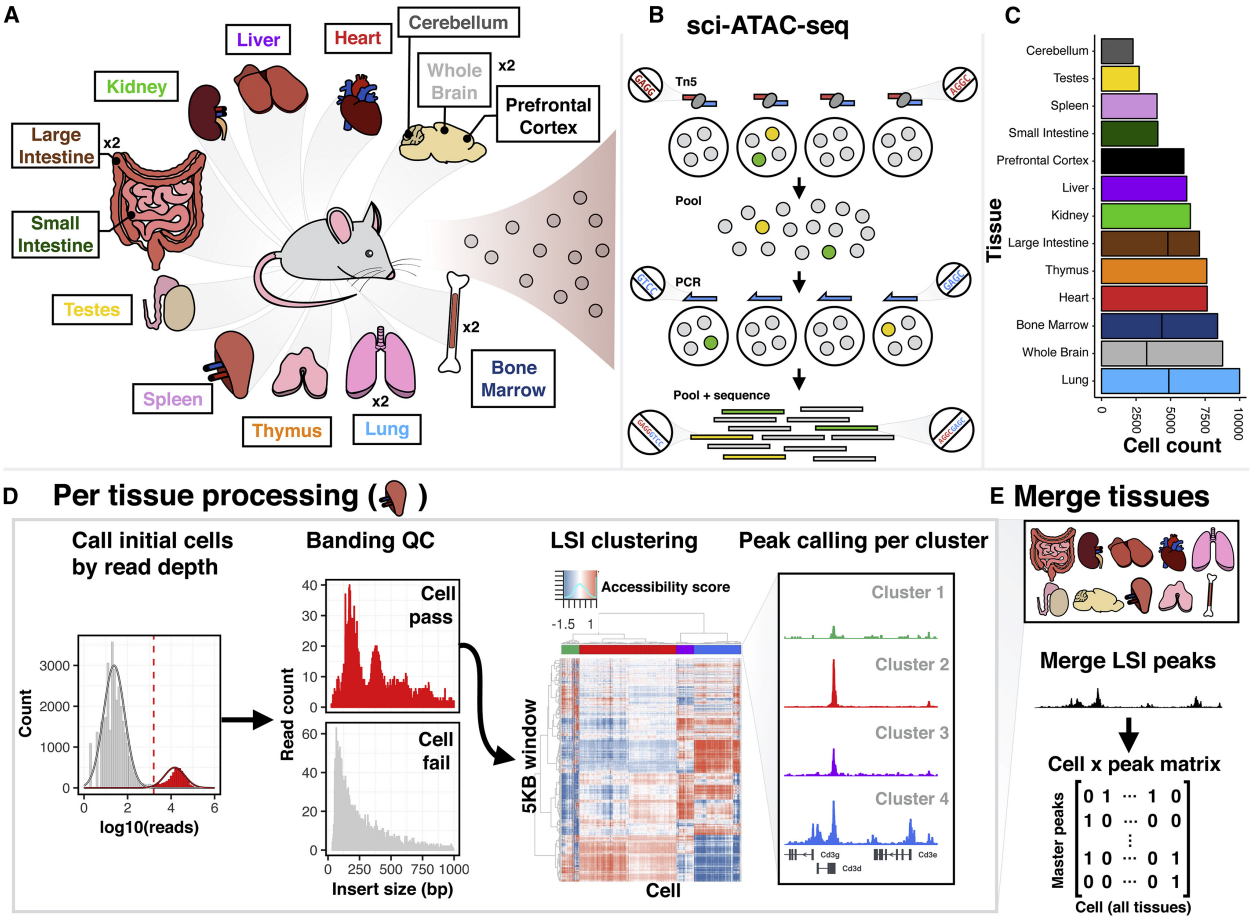


Figure 4.1: Workflow for Generating Chromatin Accessibility Profiles from Single Cells in Mice. A) Schematic of collected tissues. "×2" indicates replicated tissues. B) Schematic of sci-ATAC-seq protocol. Nuclei were barcoded in wells of a plate during Tn5 tagmentation. After pooling and splitting onto a second plate, a second barcode is introduced via PCR. Unique combinations of barcodes identify reads from single cells. C) Count of cells from each tissue passing QC. D) Example of QC steps and peak calling (data shown from spleen). Low-read-depth barcodes and barcodes lacking strong banding patterns are filtered. Cells are scored for insertions in 5 kb windows across the genome, normalized using latent semantic indexing (LSI), and clustered. Peaks are called separately on each cluster. E) Peaks from all clusters across all tissues are merged into a master peak set for a binary cell × peak matrix indicating any reads occurring in each peak for each cell.

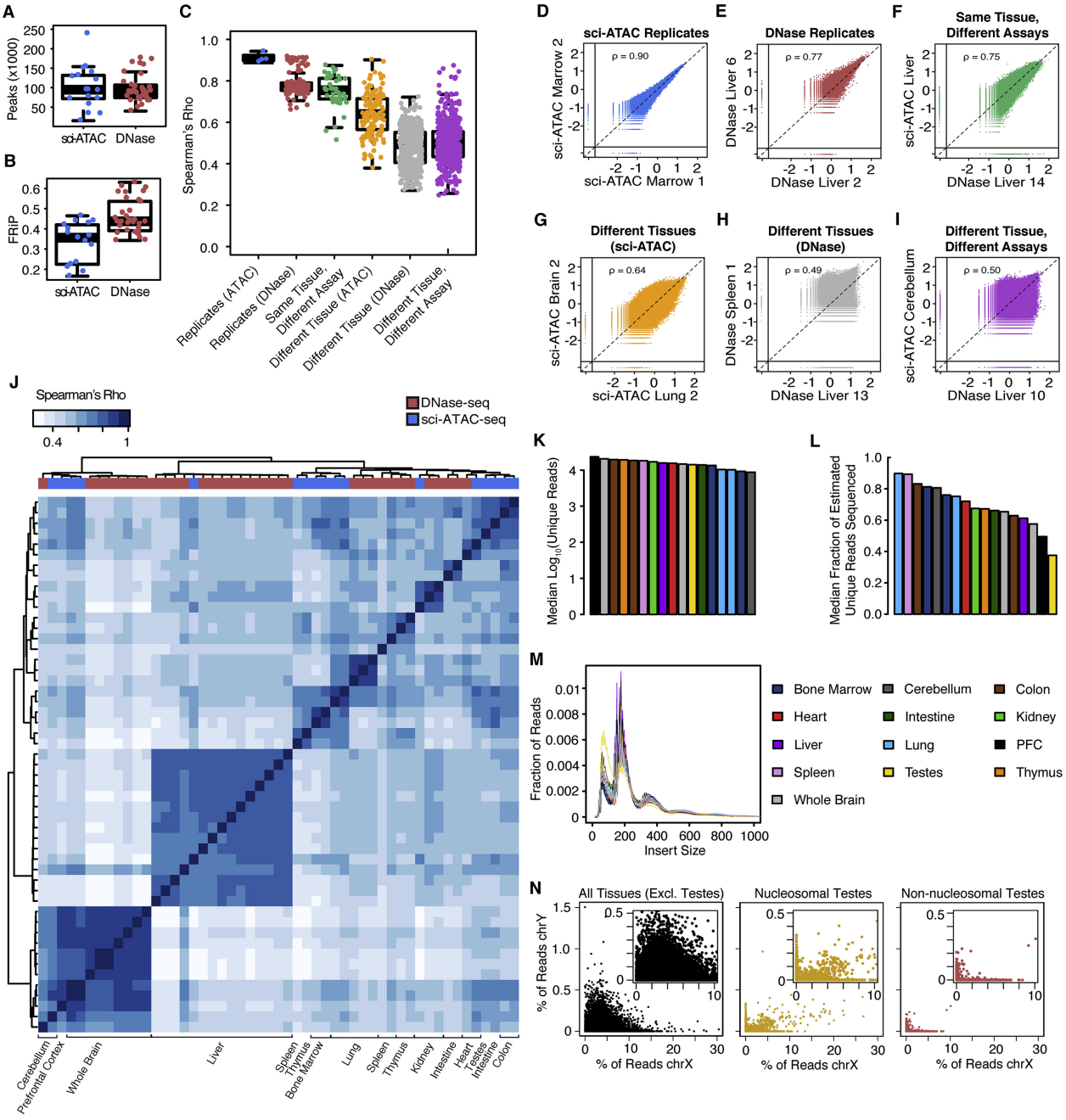


Figure 4.2: Assessing the Quality of sci-ATAC-Seq Libraries. A) Boxplot of the number of peaks of accessibility identified for sci-ATAC-seq or ENCODE DNase-seq on an overlapping set of tissues. B) Boxplot of FRiP scores for individual samples. C) Boxplot of pairwise Spearman correlation coefficients for quantitative measures of accessibility stratified by certain classes of comparison. (*legend continued on next page*)

(*continued*) D–I) Scatterplots of pairwise comparisons of accessibility (median correlation for each class shown). Log10 reads per million ("RPM") unique mapped reads in the union of all tissue-level peaks shown. Dashed line: identity line. Solid lines: distinguish sites with zero reads in one sample. J) Bi-clustered heatmap of Spearman correlation coefficients. Inverse Spearman's rho used as a distance metric, clustered with Ward's algorithm. K) Bar plot of median log10(unique reads) per cell in each tissue. L) Bar plot of the fraction of estimated unique reads that have been sequenced for each cell (implementing the same complexity algorithm as Picard - `http://broadinstitute.github.io/picard`). M) Distribution of sequenced insert sizes for each tissue. Color legend for each tissue used in panels K-M. N) Scatterplot of reads mapping to the sex chromosomes for individual cells. Insets: zoomed in view of same data. X-axes: percent of unique reads from X chromosome. Y-axes: percent of unique reads from Y chromosome. Left panel: cells from all tissues (excluding testes). Center panel: cells from the testes that passed nucleosomal banding threshold. Right panel: cells from the testes that failed nucleosomal banding threshold.

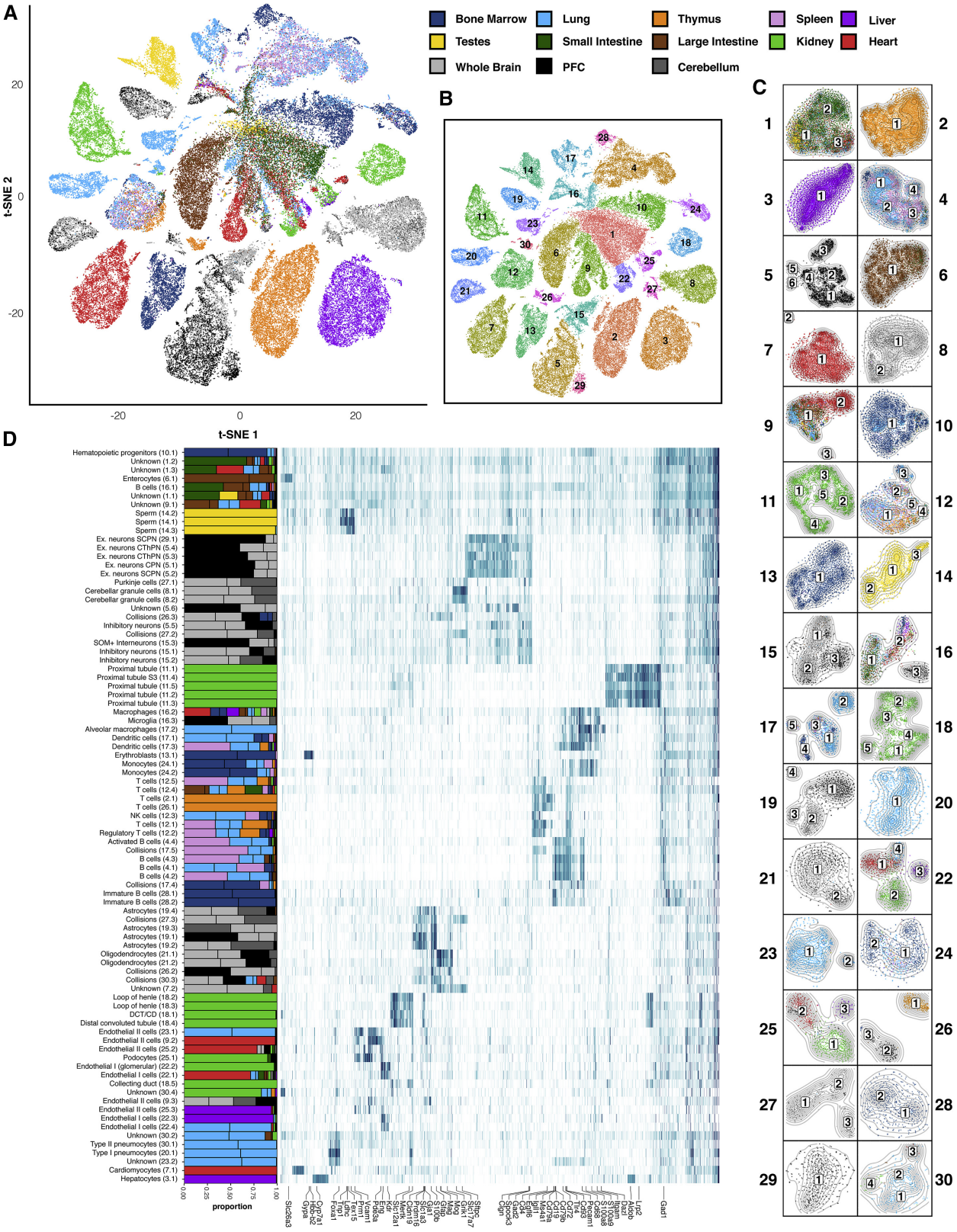
A hallmark of high-quality ATAC-seq libraries is a banded insert size distribution with peaks resulting from nucleosome protection (Figure S1M), which was apparent even in individual cells (Figure 4.2D, "banding QC" lower panel). We therefore developed a fast Fourier transform-based metric to quantify nucleosomal banding and excluded another 6,140 cells (6%) due to poor nucleosomal signal. Interestingly, 48% (2,918) of "poorly banded" cells were from testes. We speculated that these correspond to sperm or sperm precursors in which histones are replaced with protamines during maturation. Consistent with this, nearly all poorly banded testes cells (Figure 4.2N, right), as well as many of the well-banded testes cells (Figure 4.2N, middle), appear to exclusively harbor either an X or Y chromosome. The latter likely corresponds to sperm progenitors after meiosis I but prior to the histone-to-protamine transition.

4.3.1 Identifying Clusters of Cells with Similar Chromatin Landscapes

Toward identifying cell types, we first generated a master list of 436,206 accessible sites (Figure 4.1D, right panels, and STAR Methods) (Cusanovich et al., 2017) and then scored all cells for the presence of reads at these sites (Figure 4.1E). After removing poorly sampled sites and cells, we subjected 81,173 cells (see Additional Resources) to t-distributed stochastic neighbor embedding (t-SNE) (Figure 4.3A) and identified 30 major clusters of cells using Louvain clustering (Figure 4.3B). Reassuringly, some clusters were overwhelmingly derived from one tissue (e.g., 97% of cluster 7 is from heart, likely cardiomyocytes; 99% of cluster 3 is from liver, likely hepatocytes) (Figures 4.3A and 4.4A). Furthermore, most cells from some tissues appear in just one cluster (e.g., 53% of heart-derived cells are in cluster 7; 91% of liver-derived cells are in cluster 3) (Figures 4.3A and 4.4B). In contrast, other clusters include cells from many tissues (e.g., cluster 4 derives from lung [44%], spleen [44%], bone marrow [5%], large intestine [2%], and others), and some tissues are distributed across many clusters (e.g., whole brain contributes to clusters 8 [34%], 5 [17%], 15 [13%], 21 [11%], and others) (Figures 4.3A, 4.4A and B). Replicate samples of the same tissue from different mice are similarly distributed despite being processed in different batches (Figures 4.4C and D).

Because heterogeneity was apparent within many of the 30 major clusters, we adopted an iterative strategy taking cells from each major cluster and repeating t-SNE and Louvain clustering to identify subclusters (Figure 4.3C). This procedure yielded 85 distinct patterns of chromatin accessibility. Of note, 89% of the 436,206 initially identified sites were significantly differentially accessible (DA) at a false discovery rate (FDR) of 1% in at least one of these 85 cell clusters relative to a control set of 2,040 cells (120 cells randomly sampled from each of the 17 samples; see Additional Resources).

To identify DA sites at which accessibility was restricted to specific cluster(s), we adapted a metric for quantifying gene expression specificity in scRNA-seq studies (Cabili et al., 2011) to chromatin accessibility and calculated it for all 436,206 sites by all 85 clusters. We classified 39% (167,981/436,206) of accessible sites as cluster restricted (i.e., increased accessibility in a limited



(*legend on next page*)

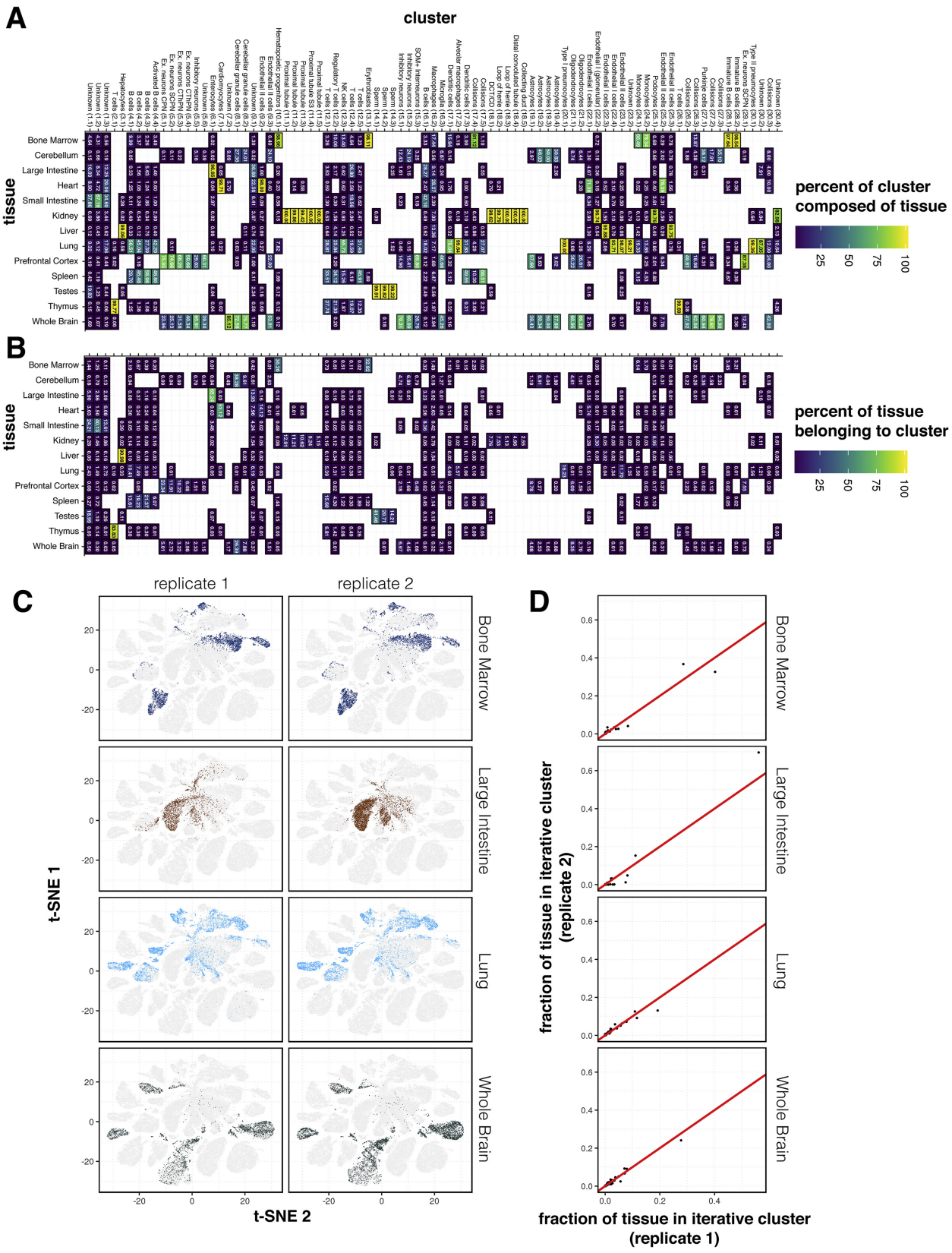
Figure 4.3: Clustering of Single-Cell Chromatin Accessibility Identifies Diverse Cell Types. A) t-SNE embedding of all cells from the dataset colored by tissue. B) Same as (A), colored/labeled according to the 30 clusters in the first round of clustering. C) Iterative t-SNE embeddings for cells from each cluster in (A) colored by tissue and labeled by their iterative cluster (85 total clusters). D) Heatmap of beta values from differential accessibility tests relative to reference cells. The numbers in parentheses correspond to the clusters in (C) (major iterative cluster). A sampling of 10,000 sites that are significantly more accessible than in the reference for at least one cluster are shown along with the promoters of relevant genes (highlighted along the bottom). The proportion of each cluster originating from each tissue is shown alongside the heatmap.

number of clusters); 55% (92,334/167,981) of these were restricted to a single cluster (Figure 4.5; see Additional Resources).

4.3.2 Assigning Cell Types to Clusters

We next sought to annotate these 85 clusters. Cell-type identification from scATAC-seq is more challenging than from scRNA-seq, largely because we have fewer guideposts in the literature. Nonetheless, many clusters were tentatively identifiable based on cluster-specific promoter accessibility of cell-type-specific genes. For example, the promoters of β-globin subunits 1 and 2 were specifically accessible in cluster 13.1 (likely erythroblasts). To incorporate information from distal regulatory sites, we applied Cicero, a method that connects distal accessible sites to promoters on the basis of scATAC-seq data (Pliner et al., 2018). Cicero reports a co-accessibility score based on how correlated two sites are in the cells comprising each cluster.

Across all 85 clusters, Cicero identified 4.4 M connections between open sites (median 39,502 connections per cluster, median number of open sites per cluster 85,731; STAR Methods and Additional Resources). These comprise a map of possible in vivo cis-regulatory interactions and include distal-to-distal (50%), distal-to-proximal (37%), and proximal-to-proximal (11%) connec-



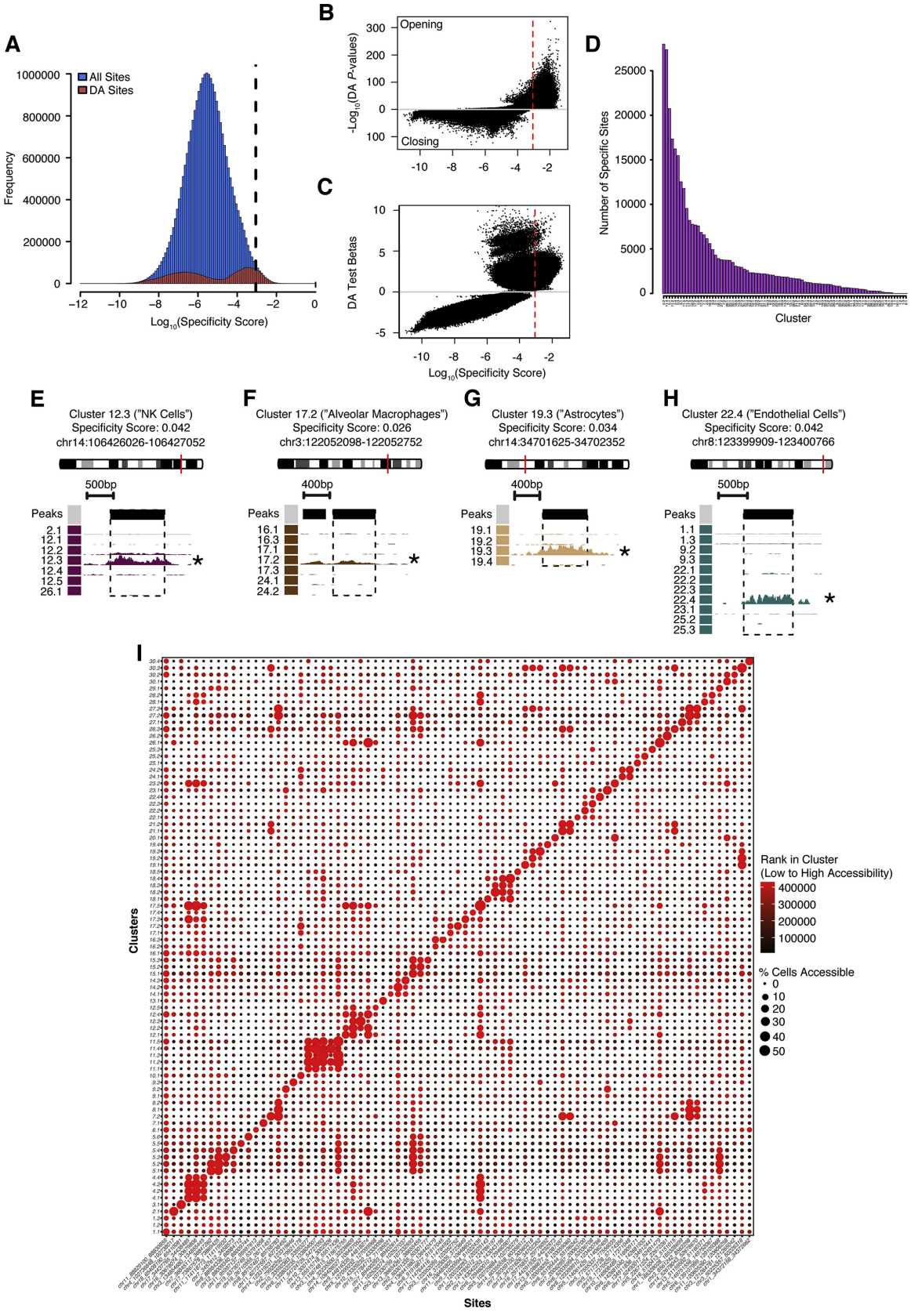
(*legend on next page*)

Figure 4.4: Chromatin States Are Reproducibly Discovered across Replicate Experiments. A) A heatmap of the percentage of each of the 85 clusters that is derived from each tissue source. B) A heatmap of the percentage of cells derived from each tissue source that belong to each cluster. Replicates are collapsed in both (A) and (B). C) t-SNE embeddings of all cells colored by replicate for the four tissues where replicate samples were collected and processed in separate batches. Replicates are plotted separately on the left and right of the plot and each tissue is included as a separate row. All cells are shown in light gray with cells from that tissue/replicate combination highlighted in color. D) Scatterplots comparing the proportion of cells from each replicate belonging to every cluster (of the 85 iterative clusters) in which 1 or more cells from either replicate is observed. The red line marks where equal proportions would lie on the plot.

tions. 64% (232,766/362,293) of distal sites open in at least one cluster were linked to one or more proximal sites (i.e., promoters). Notably, 29% (69,071/232,766) of these distal sites were linked to a promoter in only one cluster. Considering only clusters with 1+ distal-to-proximal connection for any given promoter, promoters were linked to an average of 4.9 distal sites per cluster.

Enrichments of Gene Ontology (GO) terms for genes linked to DA promoters in a given cluster were often strongly indicative of a single cell type (see Additional Resources). For example, genes with significantly open promoters in cluster 3 (likely hepatocytes) are strongly enriched for terms related to lipid metabolism (Figures 4.6B and C). We further tested for GO enrichments considering only genes linked to distal DA sites and also found them to be informative (Figures 4.6C and D). Similarly, mouse-specific annotations also implicated individual cell types. For example, considering the Mammalian Phenotype Ontology (MPO), an annotated catalog of spontaneous, induced, and genetically engineered mouse mutations (Smith et al., 2005), we found that the genes with DA promoters or linked distal sites in a given cluster were often enriched for cell-type-specific functions (see Additional Resources).

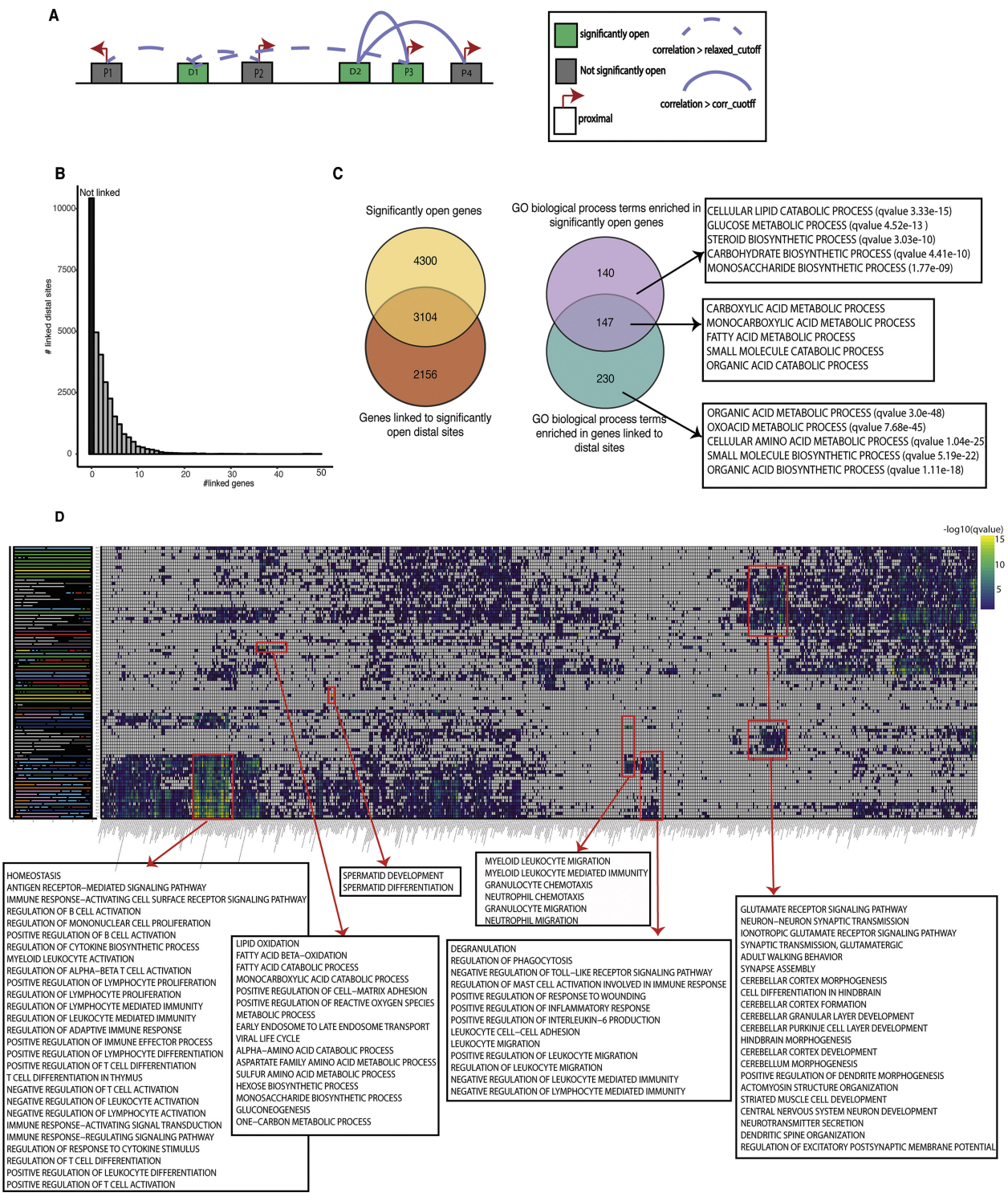
We also aggregated information across all DA sites linked to a target gene to compute a quan-



(*legend on next page*)

Figure 4.5: Specificity Scores Identify Marker Sites for Individual Cell Clusters. A) Histogram of specificity scores (see STAR Methods) for all tests (blue). Site/cluster combinations that had significant DA tests overlaid in red. Dashed line indicates threshold for calling a site "specific." B) Specificity scores (x axis) plotted against -log10(P-value) (y axis) for the differential accessibility tests. Only site/cluster tests that were significant (red bars in (A)) are shown. Significantly opening sites are plotted above the gray line and significantly closing sites are below the gray line. C) Specificity scores (x axis) plotted against beta values (y axis) for the differential accessibility test. Only site/cluster tests that were significant (red bars in (A)) are shown. y axis was truncated at 10 for visualization. D) Bar plot of the number of sites identified as "specific" for each cluster ordered from most to least. E–H) Examples of browser tracks for the top specific site for 4 different clusters. Top track (black bars plotted) below the ideogram and scale bar indicates the location of peaks of accessibility identified in each window. Windows are centered on the top site for the cluster being highlighted. Subsequent tracks (various colors) show the reads mapping to this window for various clusters. An $*$ marks the cluster for which the indicated site was specific. The other tracks presented were identified as related cell types. I) "Dotplot" of the top site (x axis) for each cluster (y axis). Clusters were ordered numerically. Sites were ordered according to which cluster for which they were the most specific site (4 clusters had no sites meeting the specificity threshold). The diameter of each dot represents the percentage of cells for which that site was accessible in the cluster. The color of the dot indicates the accessibility rank of that site relevant to other sites in that cluster ranging from black (low accessibility relative to other sites) to red (high accessibility relative to other sites).

titative "gene activity score" (see Additional Resources). We have found that these Cicero-based activity scores correlate with gene expression more faithfully than promoter accessibility alone (Pliner et al., 2018). After curating a set of marker genes from the literature corresponding to expected cell types, we estimated their activity scores in each of the 85 clusters (see Additional



(*legend on next page*)

Figure 4.6: Putative cis-Regulatory Maps Associate Chromatin States with Cellular Functions. A) Schematic of how distal sites are linked to genes. Distal sites are linked to putative target genes in each cluster using Cicero if they are co-accessible with a proximal site (within 500 bp of an annotated TSS). For linking, we used two thresholds. First, a distal site is linked to any gene if it has co-accessibility > 0.2 with the proximal site. Second, for any distal site that is not yet linked to at least 1 gene, we assign the site to the gene it is most co-accessible with provided that the co-accessibility score is > 0.1. B) Bar plot showing number of linked distal sites to genes in cluster 3.1 (hepatocytes). C) Venn diagrams showing i) the overlap between genes that were linked to distal sites using Cicero and genes with significantly open promoters, ii) the overlap between significantly enriched GO "Biological Process" (BP) terms for these two gene sets. D) Heatmap showing GO BP enrichment results (-log10(q-value)) for the union of genes linked by Cicero to distal DA sites and proximal DA sites. Terms that were enriched (q-value ¡ 0.0001, fold change Math Eq2) in at least one cluster, and at most in 50 clusters, are shown. The corresponding table of terms by enrichment values was binarized (if term was significant = 1, otherwise 0) and then clustered using hierarchical clustering. Enrichment -log10(q-values) were capped at 15 for visualization in the heatmap.

Resources). This enabled the assignment of 51 clusters to a specific cell type. However, some clusters were not positive for any of our markers, while others had high activity scores for markers associated with multiple cell types. We therefore developed a classifier trained on the accessibility profiles of marker-associated cells that allowed us to assign cell types to 12 additional clusters (STAR Methods).

We manually reviewed these automated assignments and made adjustments as deemed appropriate based on focused consideration of selected subsets of cells at either the whole tissue level (e.g., kidney) or at the level of broad cell types (e.g., all neurons). In addition, we used gene activity scores to identify genes whose activity was restricted to one or several clusters (see Additional

Resources). Both site-level and gene-level specificity scores were informative in refining our cell type assignments. For example, cluster 12 was divided into five subclusters, all designated simply as "T cells" by our automated assignments. However, both site and gene specificity scores highlighted genes with known roles in the function of regulatory T cells for cluster 12.2 and natural killer (NK) cells for cluster 12.3. In total, we assigned cell types to 69/85 clusters; based on the patterns of site usage, seven appear to be mixtures of cell types due to collisions and nine remain unknown (Figure 4.3D). A summary of all information used to assign cell type labels, including any post-hoc adjustments, is provided supplementary data files. A bi-clustered heatmap of differential accessibility in each cluster, at marker gene promoters plus a random sampling of 10,000 DA proximal and distal sites, illustrates broad patterns of similarity and dissimilarity among the 85 clusters (Figure 4.3D).

4.3.3 Single-Cell Chromatin Accessibility versus Gene Expression

A principal goal of the field is to construct a comprehensive atlas of mammalian cell types, which may require integrating atlases measuring different aspects of molecular biology (Lake et al., 2016). To that end, we sought to compare our single-cell chromatin accessibility atlas with recent single-cell transcriptional atlases of the adult mouse. We first examined the proportions of cell types observed in different tissues and found variable concordance. For example, cell-type representation in the kidney was reasonably consistent between our data and three scRNA-seq studies (Han et al., 2018; Park et al., 2018; The Tabula Muris Consortium et al., 2017) but much less concordant in the lung and brain even between two scRNA-seq studies (Figure 4.7). There are likely multiple contributors to discrepancies, including inherent biases for/against specific cell types with each protocol and data analysis choices made in each study.

We next sought to examine the similarity of cell-type annotations. As a first step, to validate the use of activity scores, we compared the average normalized activity score profiles for each sci-ATAC-seq cluster to average normalized expression profiles of matched tissues from two scRNA-seq atlases using Spearman correlation. We observed that the cell types with the highest correlation

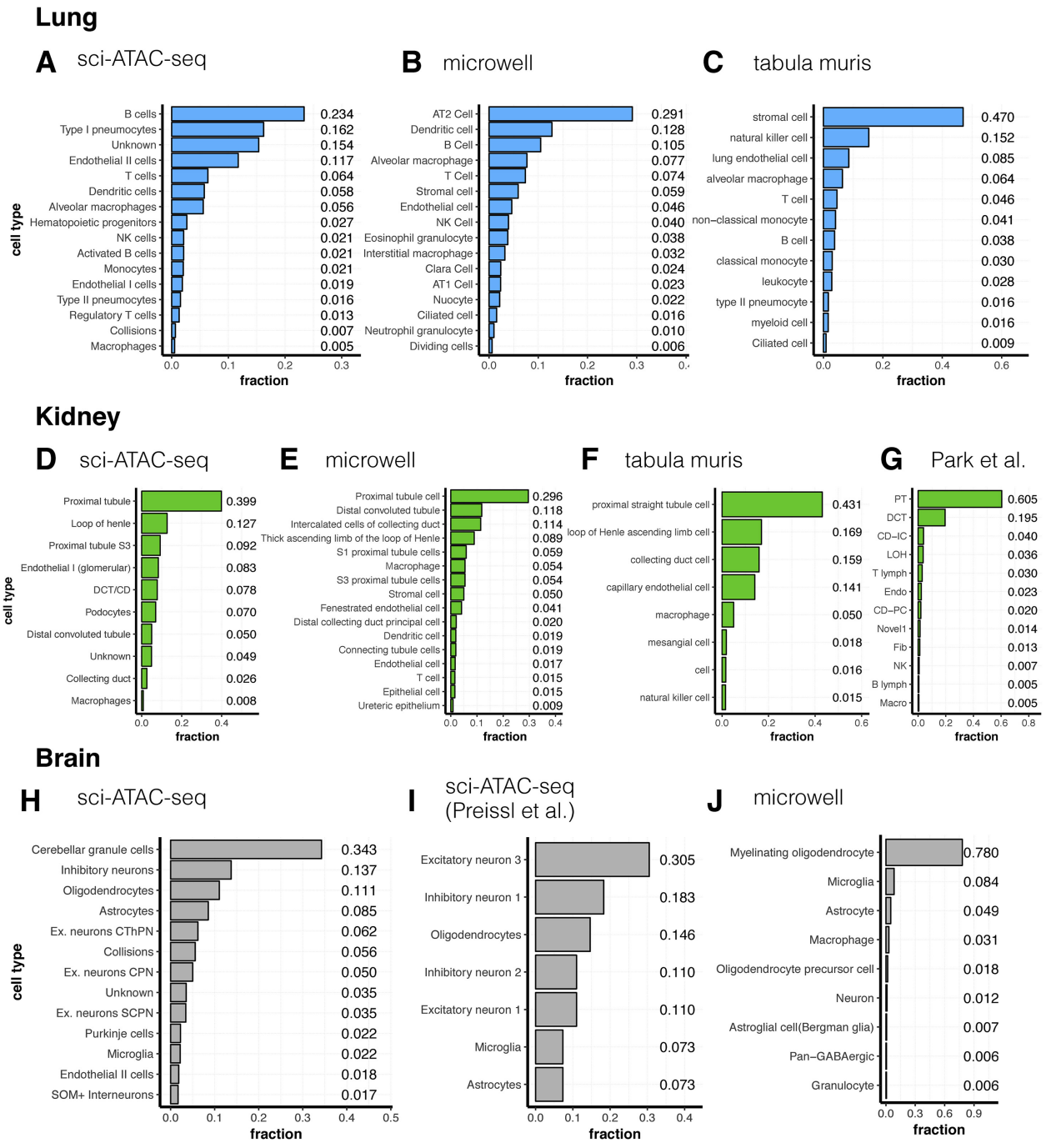


Figure 4.7: Proportions of Cell Types within Tissues Show Mixed Concordance across Available Atlases. Cell types below 0.5% frequency are excluded from all plots to facilitate visualization and labels from Han et al. (2018) are combined or shortened as deemed appropriate as in other figures (see Methods). (*legend continued on next page*)

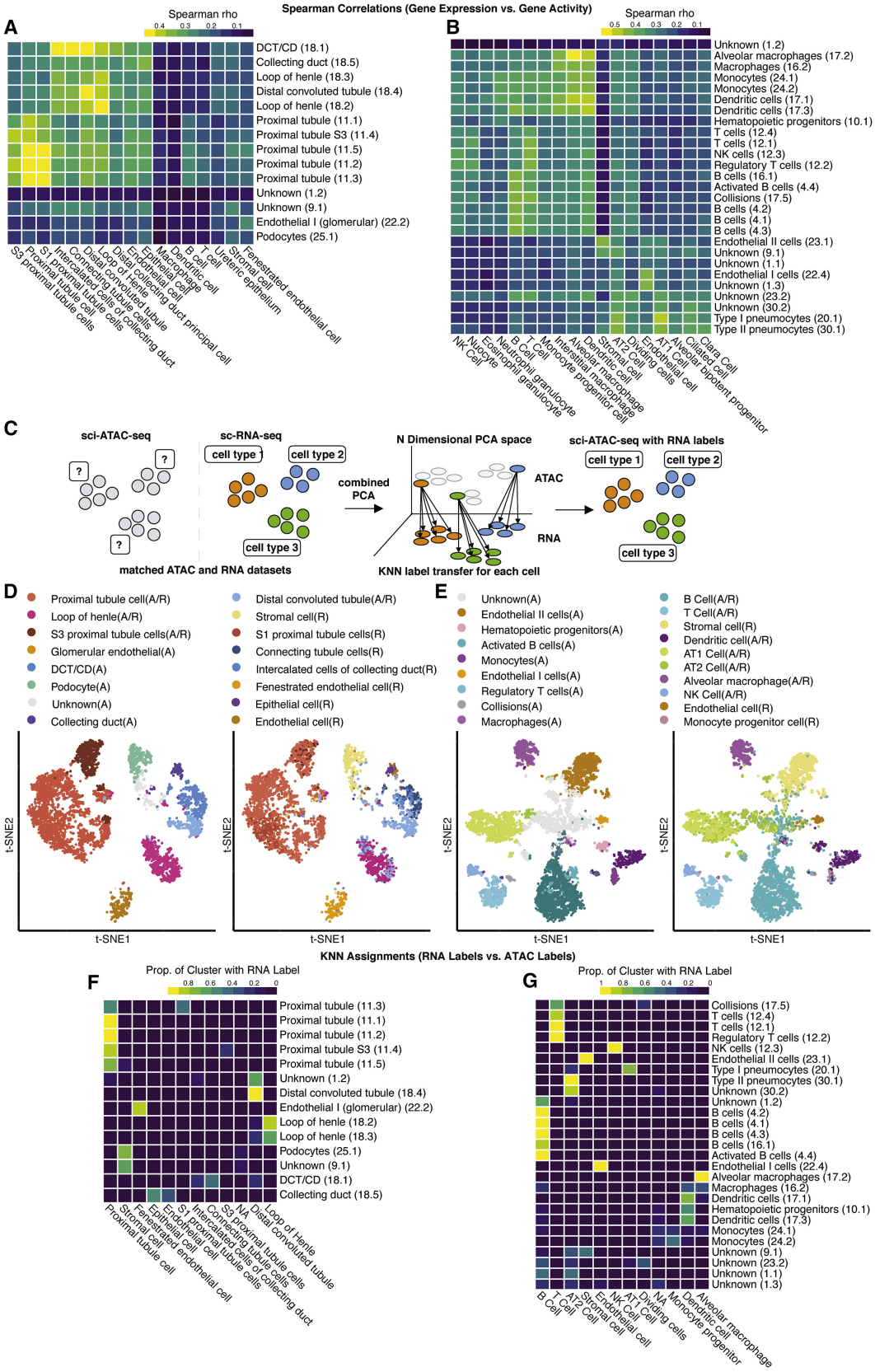
(*continued*) A–C) Comparisons of the relative proportions of cell types annotated in the Lung for our dataset, Han et al. (2018), and the The Tabula Muris Consortium et al. (2017) dataset respectively. D–G) Same for Kidney across our dataset, Han et al. (2018) , The Tabula Muris Consortium et al. (2017), and Park et al. (2018) datasets. H–J) Same for Whole Brain across our dataset, Preissl et al. (2018) (sci-ATAC-seq), and Han et al. (2018) datasets.

across datasets were concordantly annotated in the majority of cases (Figures 4.8A, 4.8B, and 4.9A and Additional Resources).

Encouraged by these results, we developed an unsupervised approach for transferring cell-type labels for individual cells from one data type to the other. After performing PCA on a combined matrix of scRNA-seq expression and sci-ATAC-seq activity scores, we used k-nearest neighbor (KNN)-based classification to transfer the most common label among their nearest scRNA-seq neighbors to sci-ATAC-seq cells. Using data and labels from two scRNA-seq atlases (Han et al., 2018; The Tabula Muris Consortium et al., 2017), we found that their cell-type assignments were largely concordant with our labels for many overlapping tissues (Figures 4.8C–G and 4.9B and Additional Resources), suggesting that relatively simple methods facilitate joint analysis of expression and chromatin accessibility data from matched tissues. However, we note two limitations of this method: (1) its performance is was much less reliable on tissues with large class imbalance (dominated by a single-cell type, e.g., thymus) and (2) it does not handle cases where a cell type appears in one dataset but not the other. Both limitations are likely addressable via optimization and adoption of a mutual nearest-neighbors approach (Haghverdi et al., 2018), respectively.

4.3.4 A Complex Sequence Grammar Underlies In Vivo Chromatin Accessibility in Cell Types

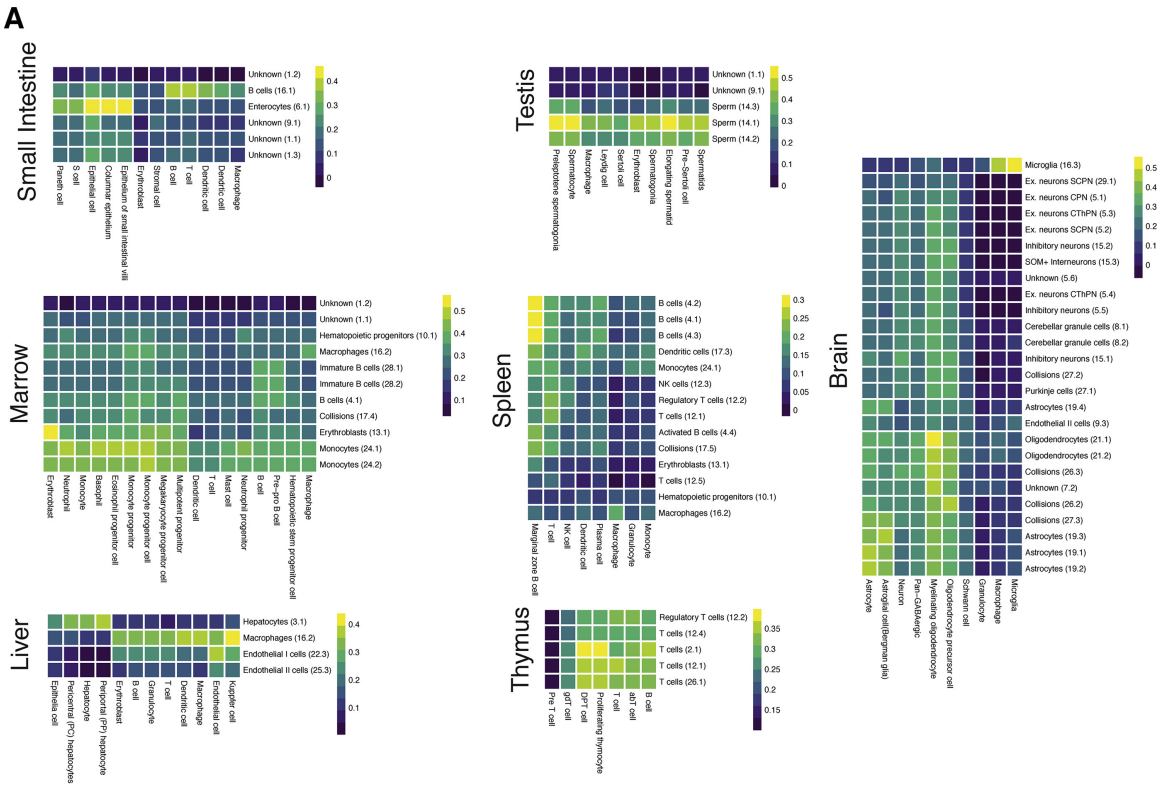
We next investigated the TF regulatory grammar that underlies in vivo chromatin accessibility in each mammalian cell type. We used Basset (Kelley et al., 2016) to train a convolutional neural network (CNN) to predict which sites are accessible in each of the 85 clusters (Figure 4.10A).



(*legend on next page*)

Figure 4.8: KNN-Based Approach Allows for Comparison of sci-ATAC-Seq and scRNA-Seq Atlases. A and B) Heatmaps of Spearman correlations between average normalized expression/activity score profiles for groups defined in Han et al. (2018) (scRNA-seq; x axis) and our dataset (y axis) for kidney and lung, respectively. C) Schematic of KNN-based approach for transferring labels from scRNA-seq data to sci-ATAC-seq data cell by cell in principle-component analysis (PCA) embedding (see STAR Methods). D) t-SNE embedding colored by labels made on sci-ATAC-seq data alone (left) and labels derived from KNN using Han et al. (2018) kidney data (right). Labels are annotated with "(A)" if term was used in this ATAC study, "(R)" if used in Han et. al 2018, and "(A/R)" if used in both (there are some discrepancies in the labels used by the two studies). Similar colors indicate similar annotations. E) Same as (D) for lung. F and G) Matrix of the proportions of each sci-ATAC-seq category that maps to each category from Han et al. (2018). Note that some labels from Han et al. (2018) were shortened (see STAR Methods). Only cell types at or above a 0.5% frequency are shown for each study. scRNA-seq cells with fewer than 600 unique molecular identifiers (UMIs) across genes common to both datasets and sci-ATAC-seq cells with fewer than 1,800 non-zero values in the master peak set were excluded to improve KNN performance (also used to compute correlations with little impact on results).

The CNN learned to discriminate cells in each cluster from all other cells on the basis of the sequence content of accessible sites. After the CNN was trained, we annotated the 600 first-layer convolution nodes, each comprising a weighted matrix of sequence features ("filters") similar to a TF motif position weight matrix. The motif analysis tool TomTom (Bailey et al., 2009) assigned 278/600 filters to known motifs (Kulakovskiy et al., 2016). To assess which filters were most relevant to individual cell types, we removed them one by one and measured the drop in predictive performance (Figure 4.10B). We also developed an aggregate score for each filter quantifying the extent to which it is represented in accessible sites observed in individual cells (see Additional Resources). With this framework, we were able to extend our approach from clusters of cells to



(*legend on next page*)

Figure 4.9: Comparing scRNA-seq and sci-ATAC-seq Datasets. (A) Heatmaps of spearman correlations between average normalized expression / activity score profiles for groups defined in Han et al. (2018) (scRNA-seq; x axis) and our dataset (y axis) for matched tissues, related to Figure 4.8. B) Matrices of the proportions of each sci-ATAC-seq category (y axis) that maps to each category from Han et al. (2018) (x axis) using our KNN based approach presented in Figure 4.8 for matched tissues, related to Figure 4.8. See Figure 4.8 and Methods for details, particularly regarding derivation of the labels used for Han et al. (2018) and filtering of low abundance labels.

single cells and identify plausible motifs even for clusters with very few DA sites. For example, cluster 12.5, the cluster with the fewest DA sites (359), ranks a filter matching the GATA motif as highly influential, consistent with the role of Gata3 in T cell development (Hosoya et al., 2010) (Figure 4.10C, top panel). The filter best matching the MEF2 motif is highly influential not just for sites accessible in cardiomyocytes, but also in neurons and hematopoietic progenitors (Canté-Barrett et al., 2014; Rashid et al., 2014) (Figure 4.10C, second panel). Likewise, the classification accuracy of hepatocytes, enterocytes, and kidney epithelia are strongly influenced by the filter matching the PPAR motif (Figure 4.10C, third panel). Finally, this framework can be used to identify novel motifs (Figure 4.10C, bottom panel; broadly influential with the exception of non-cerebellar neurons, sperm, and a few others).

4.3.5 Specialization of Cell Types Distributed across Tissues

An open question is whether cell types distributed throughout the body exhibit tissue-specific chromatin architecture. We first investigated this question by focusing on endothelial cells. We grouped cells from clusters labeled as endothelial and re-analyzed them, resulting in nine distinct clusters (Figure 4.11A), each of which exhibited tissue specificity (9/9 with ¿50% and 5/9 with ¿90% of cells from one tissue; Figure 4.11B). Interestingly, endothelial cells from brain and kidney largely fell into their own clusters, while those from lung and liver were each split between two clusters,

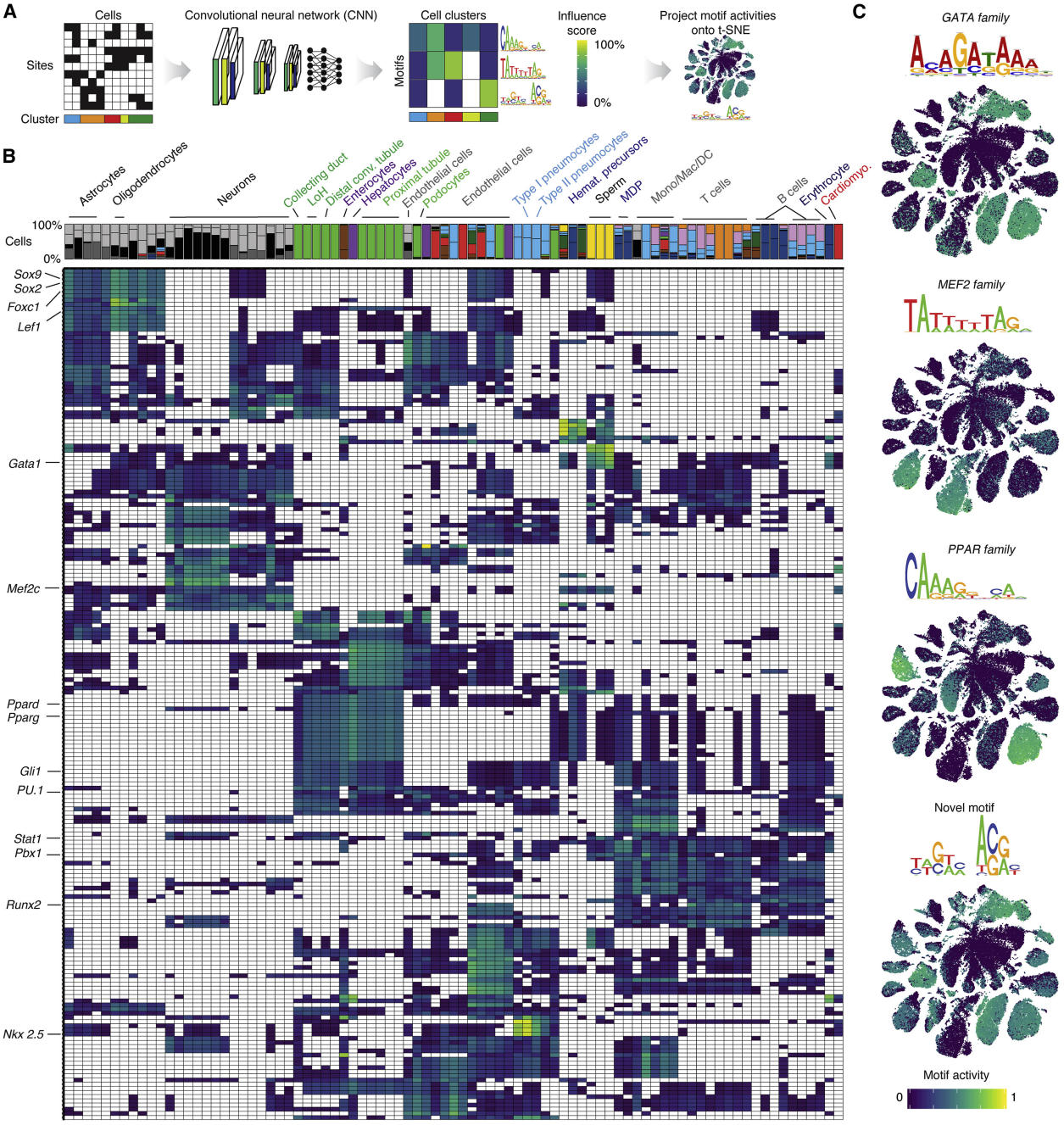


Figure 4.10: Cell-Type-Specific Chromatin Accessibility Is Associated with a Complex Sequence Grammar. A) Schematic of steps for finding motifs specific to clusters by training a CNN and postprocessing of the first layer convolution nodes (called "filters"). DA sites from each cell cluster were fed into the Basset framework. Filters were annotated by similarity to known motifs, and their influence on classification was evaluated. Usage of motifs was projected onto the t-SNE of all cells. (*legend continued on next page*)

(*continued*) B) Heatmap showing normalized influence of motif-annotated filters on classification. Only positive influence scores colored in heatmap. Barplot on top indicates proportion of cells in each cluster from each tissue. Selected filters matching known motifs are highlighted on left. C) t-SNEs of motif activity for selected filters.

and those from heart were split between three clusters. We found accessibility-based gene activity scores for Flt4 (Figure 4.11C) and EphB4 (data not shown), markers of venous endothelium, to be elevated in one set of clusters, while gene activity scores for Heyl and other genes downstream of Notch signaling, which plays a crucial role in specifying arterial/venous cell fate, were elevated in the remaining clusters (Figure 4.11C). These patterns suggest these groups may correspond to venous and arterial endothelium, respectively, at least for the heart, liver, and lung. An alternative interpretation is that this dichotomy may instead reflect capillary versus other endothelial cell types, as studies have suggested that Flt4 and EphB4 may mark capillaries in addition to venous endothelium in the adult (Partanen et al., 2000; Taylor et al., 2007). Although definitive annotation will require further work, these data reveal that endothelial cells have specialized patterns of chromatin accessibility within and between tissues.

As a second example of tissue specialization, we focused on monocytes, macrophages, and dendritic cells (DCs), which are also broadly distributed. We grouped and re-analyzed cells with corresponding labels, resulting in six distinct clusters (Figures 4.11D and E). Several of these were readily identifiable by marker genes (Figure 4.11F). Cluster 3 exclusively derives from lung and has a high gene activity score for Pparg and likely corresponds to alveolar macrophages. Cluster 6 exclusively derives from brain tissues and has a high gene activity score for Sall1, a known marker of microglia (Lavin et al., 2014). The remaining clusters derived from bone marrow, lung, heart, liver, and spleen to varying degrees. Cluster 1 was mostly from marrow and had elevated chromatin accessibility around Cd24a, Vegfa, and Cd62, which mark monocytes. Cluster 2 derived from heart, liver, and kidney and had elevated chromatin accessibility near macrophage-

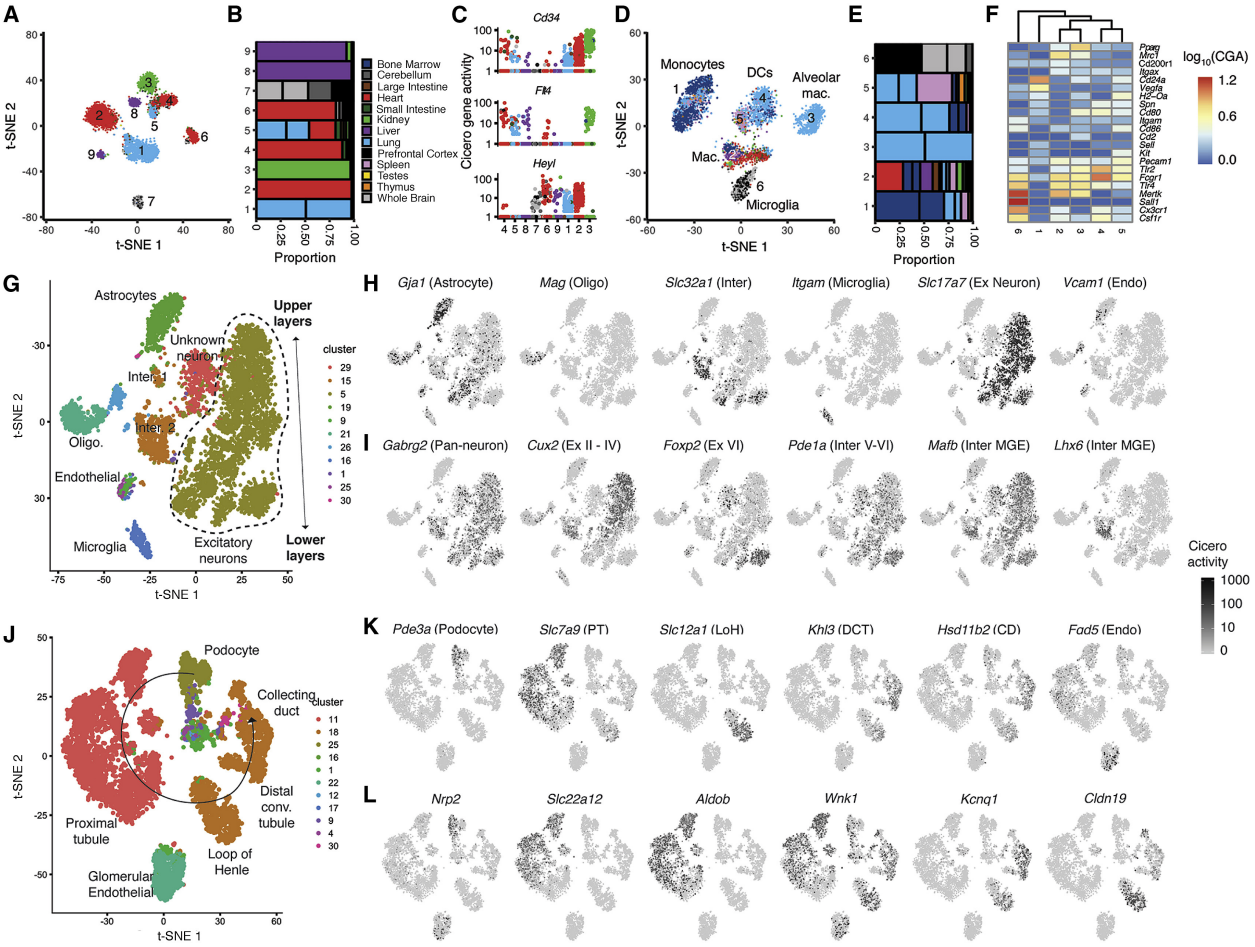


Figure 4.11: Chromatin Structure Reflects Cellular Specialization and Tissue Spatial Architecture. A) t-SNE of endothelial cells (clusters 9.2–3, 22.1–4, 23.1, and 25.2–3; colored by tissue). Numbers indicate resulting subclusters. B) Proportion of each endothelial subcluster contributed by each tissue. C) Cicero gene activity scores for selected marker genes. D) Same as (A), but for macrophages, monocytes, and dendritic cells (clusters 16.2–3, 17.1–3, and 24.1–2). E) Same as (B), but for clusters in (D). F) Heatmap showing Cicero gene activity scores for selected marker genes. G) Cells from the PFC visualized by t-SNE and colored by major cluster from Figure 4.3B. H) Cicero gene activity scores for selected markers of cell types expected in the PFC. Oligodendrocytes (Oligo), interneurons (Inter), excitatory neuron (Ex neuron), endothelial (Endo). I) Cicero gene activity scores for selected marker genes for different cortical layers (Lake et al., 2016). Ex II–IV, excitatory neurons layers II–IV; Ex VI, excitatory neurons layer VI; Inter V–VI, interneurons layer V–VI; Inter MGE, interneurons of the medial ganglionic eminence. (*legend continued on next page*)

(*continued*) J) Same as (G), but for kidney. K) Cicero gene activity scores for GWAS disease genes with restricted expression patterns (Park et al., 2018). PT, proximal tubule; LoH, loop of Henle; DCT, distal convoluted tubule; CD, collecting duct. L) Cicero gene activity scores for additional GWAS disease genes from Park et al. (2018).

associated genes (e.g., Mertk). Overall, these data demonstrate that monocytes, macrophages, and DCs adopt one of several stereotyped chromatin profiles. However, in contrast with endothelial cells, which tend to have tissue-specific patterns of chromatin accessibility, putative monocytes and macrophages appear to adopt configurations of chromatin accessibility that are more closely shared across tissues.

4.3.6 Heterogeneity in Chromatin Accessibility Can Reflect Spatial Architecture

To investigate heterogeneity in chromatin accessibility across cells classified as neurons, we reanalyzed cells from the prefrontal cortex (PFC; Figure 4.11G). As expected, excitatory neurons and interneurons were clearly segregated from glial cells, microglia, and endothelial cells (Figure 5H). However, there remained striking heterogeneity within excitatory neurons, potentially reflecting differential expression and methylation in different layers of the PFC (Figure 4.11G) (Lake et al., 2016; Luo et al., 2017). For example, regulatory elements linked to Cux2, highly expressed only in layers II–IV, and Foxp2, highly expressed in layer VI, are accessible in cells at the "top" and "bottom" of the excitatory neuron cluster, respectively (Figure 4.11I). Of the two interneuron clusters, both show accessible chromatin surrounding the interneuron marker Slc32a1, but only one shows accessible chromatin around MafB and Lhx5, specifically expressed in the medial ganglionic eminence. Overall, these observations are consistent with chromatin accessibility within a cell type varying in relation to anatomical coordinates.

We performed a similar analysis of the kidney (Figure 4.11J). Recent scRNA-seq experiments characterized differential expression across functional segments of the nephron, likely consequent

to their specialized roles in filtration (Der et al., 2017; Park et al., 2018). Chromatin near genes reported by Park et al. (2018) to be expressed in a single renal cell type tended to be accessible in only one cluster (Figure 4.11K). Moreover, the t-SNE organized our cells into a pattern reminiscent of a glomerular-collecting duct axis and similarly to the spatial layout that Der et al. (2017) observed in their scRNA-seq data, suggesting that like the neurons, kidney tubule cells' chromatin accessibility varies in relation to the tissue's spatial architecture (Figures 4.11J–L).

4.3.7 Chromatin Accessibility Dynamics during Hematopoiesis

We next examined chromatin accessibility in the bone marrow, the site of adult hematopoiesis. Although t-SNE resolved several subpopulations (Figure 4.12A), a few clusters were large and did not cleanly separate cells by accessibility at genes expressed in a mutually exclusive manner in differentiated cells. We reasoned that, similar to RNA-seq, differentiating blood cells might be organized along a continuous "trajectory" of chromatin accessibility states. We therefore applied Monocle 2, which can pseudo-temporally order cells based on chromatin accessibility (Pliner et al., 2018) to the marrow, resulting in a tree-like trajectory with a prominent "root" and five major branches (F1–F5) (Figure 4.12B).

To explore which parts of this tree correspond to various stages of blood development, we projected accessibility at previously defined sets of hematopoietic enhancers (Lara-Astiaso et al., 2014) onto the tree (Figure 4.12B). Enhancers specific to erythroid or lymphoid cells were more accessible on branches F4 and F2, respectively. Myeloid-specific enhancers were more accessible on F5 and the two small branches (F1 and F3) and modestly accessible on the root. We also examined gene activity scores for lineage markers along each branch. Gene activity scores for lineage-specific markers (Cd3e, Cd19, Hbb-b1, and Cd11b/Itgam) were at or near zero on the root, but each rose sharply on one of the five branches (Figure 4.12C). In contrast, Cd34, a marker of multipotent hematopoietic progenitors, was highly active on the root but decreased to near zero at the termini of all branches except F1. These observations are broadly consistent with specification of hematopoietic progenitors into B cells (F2), T cells (F3), erythrocytes (F4), and monocytes (F5).

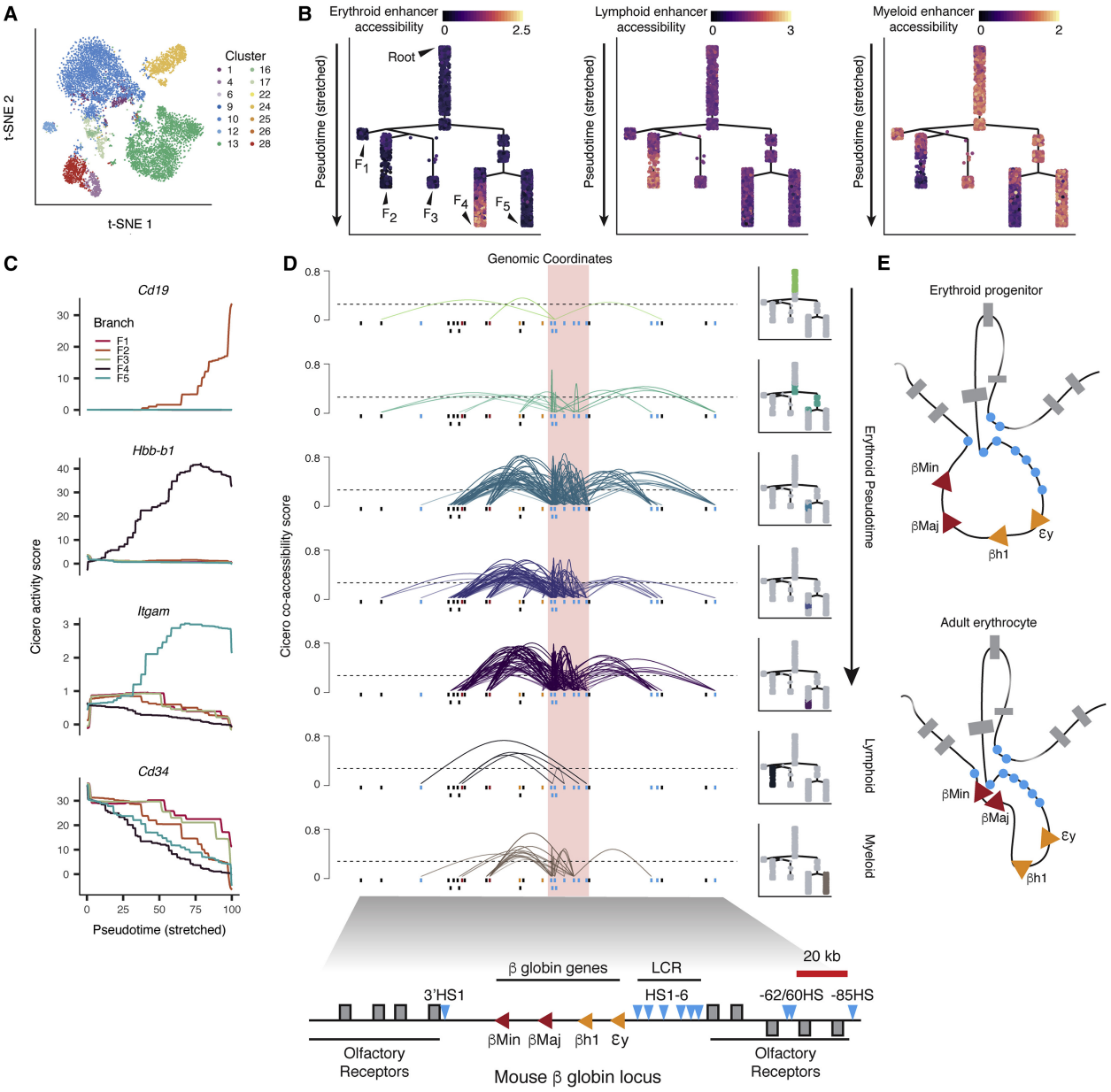


Figure 4.12: Chromatin Accessibility Dynamics during Hematopoiesis. A) t-SNE of bone marrow cells colored by major cluster from Figure 4.3B. B) Branched hematopoietic trajectory colored by accessibility of lineage-restricted enhancers (Lara-Astiaso et al., 2014). Color values represent normalized mean accessibility of peaks overlapping known enhancers (top: erythroid and erythroid progenitor, middle: lymphoid and lymphoid progenitor, bottom: myeloid and myeloid progenitor). (*legend continued on next page*)

(*continued*) C) Cicero gene activity scores of selected marker genes (Cd19, Hbb-b1, Itgam, and Cd34 for B cells, erythroid, myeloid, and hematopoietic stem cells, respectively) across pseudo-time in each branch. Each line includes cells from the root to the named branch (from B). Activity scores are plotted as a moving average over pseudo-time (percent of total distance from the root). D) Cicero co-accessibility at the β-globin LCR along erythroid differentiation (roughly equal-size groups). Cells used to generate each plot are highlighted (right). Lymphoid and myeloid plots are included for comparison. Boxes below each track indicate sci-ATAC-seq peaks (colored by overlap with elements in the β-globin locus diagrams below and in E). Arcs connecting peaks represent co-accessibility (height indicates strength of co-accessibility). Only connections originating in the LCR with co-accessibility above 0.25 (dashed line) are shown (LCR is the red highlighted region). E) Model of the β-globin locus adapted from (Noordermeer and de Laat, 2008).

We note, however, that although we would expect ˜37% of cells from marrow to be neutrophils (Yang et al., 2013), we were unable to identify any cluster or branch of cells with a consistent activity score for neutrophil markers (e.g., Elane). Neutrophil nuclei are more fragile than other cell types (Olins et al., 2008) and may not have survived fluorescence-activated cell sorting (FACS), or possibly, they are present in our dataset, but we are simply failing to identify them.

We next visualized Cicero connections for cells in different regions along the trajectory of erythropoiesis (F4). We identified sci-ATAC-seq peaks corresponding to the six hypersensitive sites (HSs) in the β-globin locus control region (LCR; HS1-6) along with several others known to play a role in establishing the 3D chromatin conformation critical for developmental control of β-globin expression (Dostie et al., 2006). In the erythroblasts of adult mice, the LCR is positioned close to β-globin subunits Hbb-b1 (βMaj) and Hbb-b2 (βMin), while during development, these genes are looped away from the LCR, which contacts subunits $\beta h1$ or εy instead (Noordermeer and de Laat, 2008; Tolhuis et al., 2002). Consistent with this, in cells at the root of the tree, Cicero reported modest co-accessibility between elements of the LCR and the more distal flanking

noncoding elements, as well as limited linkages between noncoding elements and the adult globin genes (Figure 4.12D). Cicero did not link the fetal and embryonic globins to the LCR, as expected. At intermediate stages through to the terminus of the erythroid branch, the Cicero maps have increasingly strong linkage of the adult globin genes and the LCR, the downstream 3'HS1, and both the -62/60 and -85 upstream HS (Figure 4.12D). In contrast, we observe only light links between the LCR and the other distal sites or the globin genes on the lymphoid and myeloid lineages, confirming that the robust association of the globin LCR with its targets is specific to the erythroid lineage.

4.3.8 Implicating Cell Types in Common Human Traits and Diseases

A major fraction of heritability for common human traits and diseases, as measured by genome-wide association studies (GWASs), partitions to distal regulatory elements, which are often cell-type specific. Consequently, much work has gone into intersecting GWAS signals with bulk DNase hypersensitivity data (and other epigenetic features), with the goal of systematically linking particular diseases to the dysfunction of specific tissues (Finucane et al., 2015; Maurano et al., 2012; Pickrell, 2014). However, the resolution of such studies is markedly limited by cell-type heterogeneity. Given the degree of conservation of chromatin accessibility between mice and humans (Vierstra et al., 2014), we wondered if we could use our data to better understand the cell-type-specific effects of genetic variation underlying complex human traits regardless of the differences between species. Therefore, despite the fact that our data were generated on mouse tissues, we sought to apply state-of-the-art methods for detecting cell-type-specific enrichment of human heritability (Finucane et al., 2015).

To do so, we quantified the enrichment of heritability for human traits within DA peaks for each of our 85 clusters using partitioned linkage disequilibrium (LD) score regression (LDSC) (Finucane et al., 2015). After lifting over human SNPs to orthologous coordinates in the mouse genome (Kuhn et al., 2013), we calculated the enrichment of heritability for 32 phenotypes across the DA peaks obtained for each of our 85 clusters (Figure 4.13A and Additional Resources). 55 of the 85

cell types had an enrichment for at least one phenotype, while 28 of 32 phenotypes were enriched for at least one cell type. As a broad trend, we observed a strong enrichment of heritability for autoimmune diseases such as lupus, celiac disease, and Crohn's disease in clusters corresponding to leukocytes, whereas for neurological traits such as bipolar disorder, educational attainment, and schizophrenia, the enrichments occurred in neuronal cell types (Figure 4.13B, bottom panel). Notably, most of these enrichments were not apparent in the peaks called from bulk tissues (Figure 4.13B, top panel), demonstrating the value of cell types defined by single-cell chromatin accessibility data. Many enrichments were consistent with expectation. For example, the strongest enrichments of heritability for low-density lipoprotein (LDL) cholesterol, high-density lipoprotein (HDL) cholesterol, and triglycerides are in hepatocytes, although interestingly, LDL cholesterol was also significant in the kidney epithelium of the loop of Henle (Figure 4.13B, bottom panel). Likewise, the strongest enrichment of heritability for immunoglobin A (IgA) deficiency are in clusters of T cells (Figure 4.13C). These signals can also lead to refined understandings of the importance of subtypes of cells. As an example of this trend, although enrichments of heritability for bipolar disorder are observed for multiple neuronal clusters, the strongest enrichments involve excitatory neurons (Figure 4.13D). In contrast, heritability for Alzheimer's disease is not enriched in any class of neurons. Instead, its strongest enrichment is found in a cluster of microglia (Figure 4.13E; also corresponds to cluster 6 in Figure 4.13D).

To expand our analysis to a larger set of traits, we downloaded summary statistics from `http://nealelab.github.io/UKBB_ldsc/`for GWASs for 2,419 traits in over 300,000 individuals from the UK Biobank (Bycroft et al., 2017). Focusing on 405 traits with an effective sample size of geq5,000 and estimated heritability of geq0.01, we observed significant enrichment of heritability in 273 traits in at least one cell type, while 74 of the 85 cell types exhibit enriched heritability for at least one trait (see Additional Resources). While the same broad trends as described above are also seen here for autoimmune and neurological traits (Figures 4.14A and B), the much larger number of traits measured by the UK Biobank reveals additional trends. For example, many measures of body size and composition (e.g., body mass index) are also associated with cell types in the brain (Figure 4.14B). Additionally, specific subsets of T cells (12.1, 12.2) are more associated with

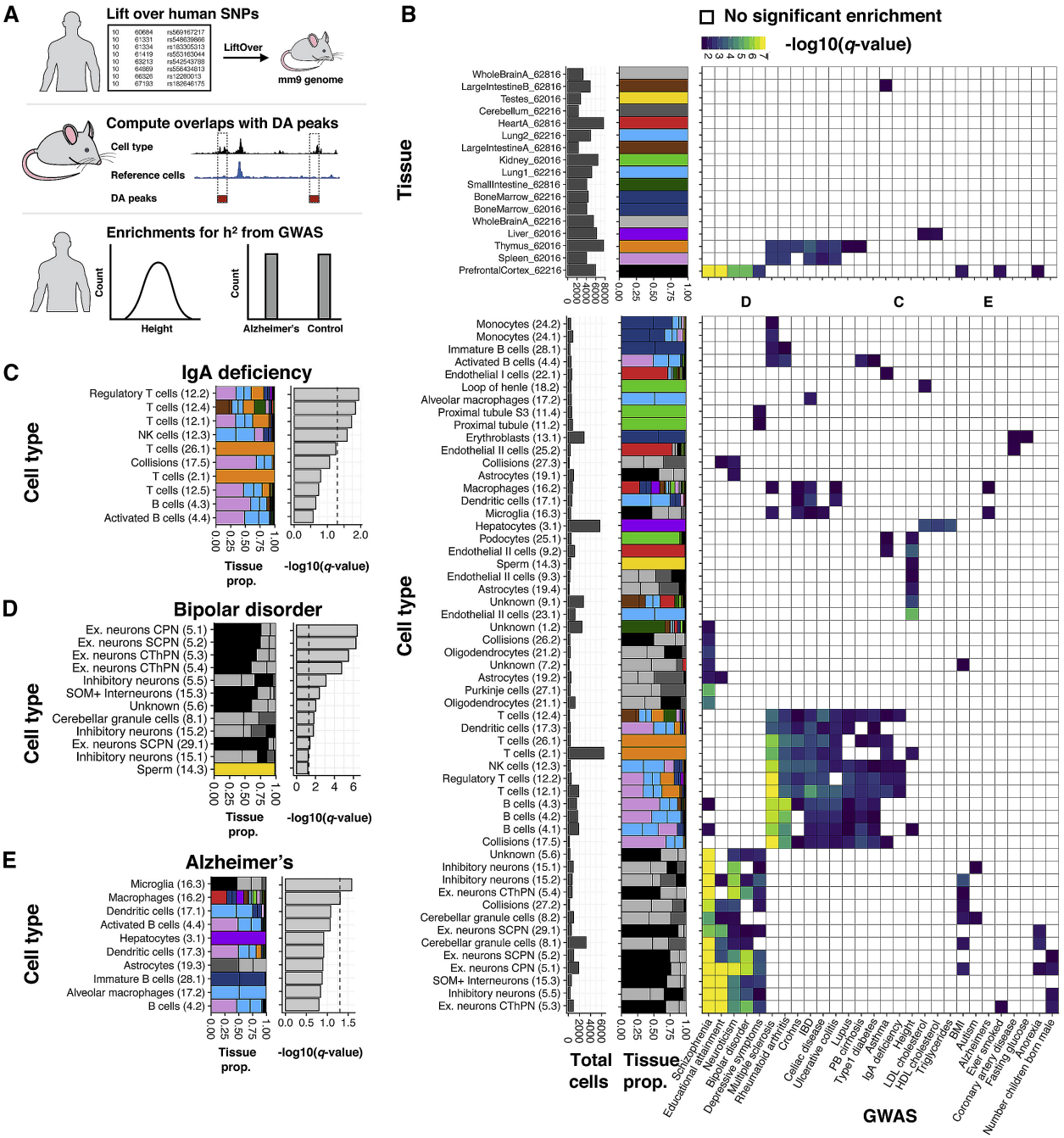
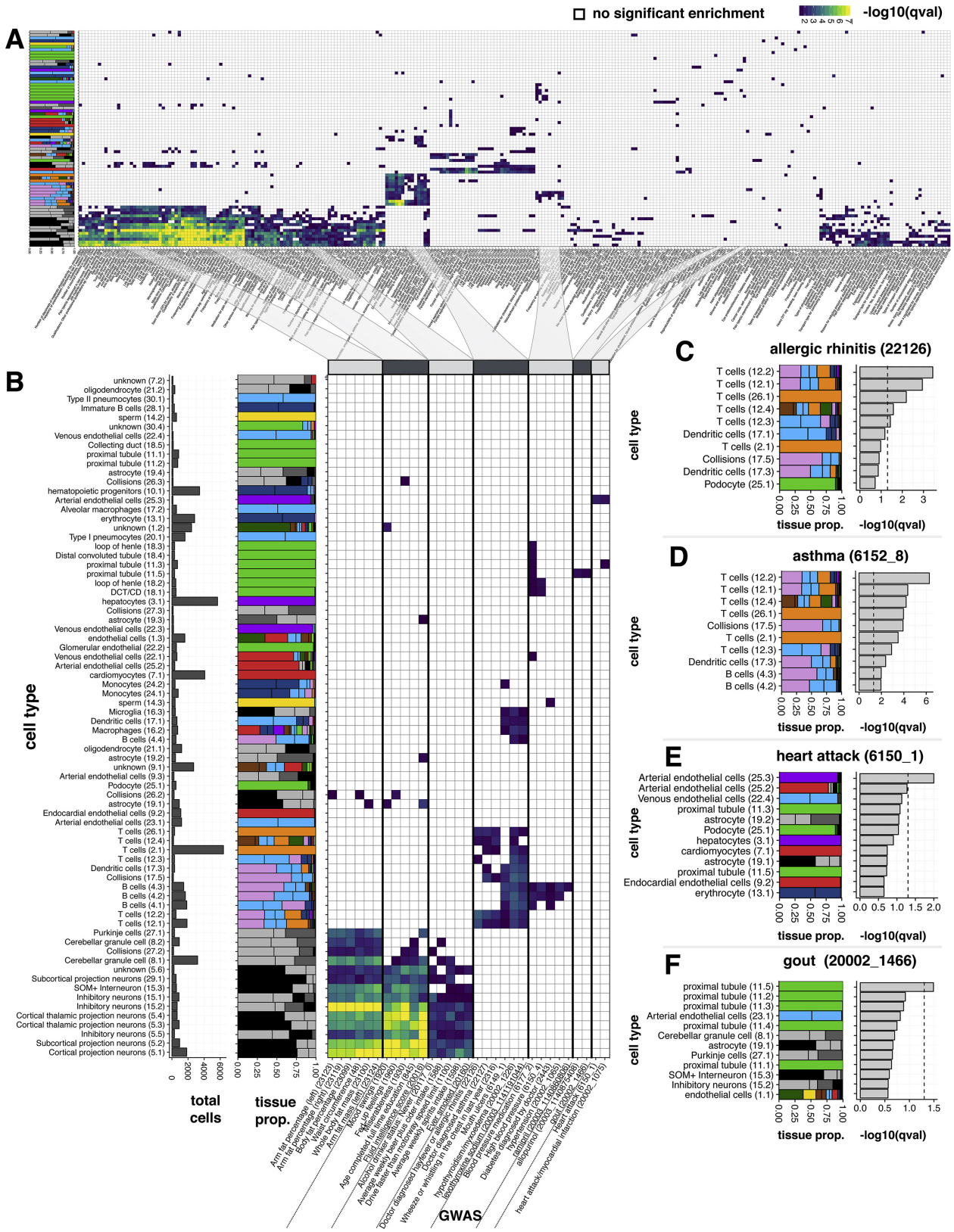


Figure 4.13: Mouse Chromatin Profiles Are Associated with Heritable Human Traits. A) Schematic of LDSC analysis workflow. Human SNPs are lifted to the mouse genome, annotated with respect to overlapping DA peaks, and used as input to LDSC. B) Heatmaps of -log10(q value of enrichment). (*legend continued on next page*)

(*continued*) Trait/cluster pairs with no significant enrichment are white. Plots to the left of each heatmap indicate the number of cells in each cluster and proportion of cells from each tissue. Upper panel shows results when using peaks called on bulk tissues. Bottom panel shows results when using peaks that are positively DA for each of the 85 iterative clusters. Letters above the lower heatmap indicate the columns for traits highlighted in C–E. C–E) Examples of individual traits (columns) from the lower heatmap in (B). (C) IgA deficiency. (D) Bipolar disorder. (E) Alzheimer's disease.Within each panel, the following are shown: the proportion of each tissue composing that cluster and -log10(q value of the enrichment). Clusters are sorted by q value, and the dotted line indicates a q value of 0.05.

asthma and allergic rhinitis than other cell types, including other T cell clusters (Figures 4.14C and D). At a more granular level, heart attacks are associated with endothelial cells from the liver (25.3), but not from other endothelial clusters (Figure S7E), while gout is associated with kidney proximal tubule cells (Figure 4.14F). The framework that we demonstrate here can be readily applied to single-cell chromatin accessibility data collected from any human or mouse tissue and any heritable trait.



(*legend on next page*)

Figure 4.14: Demonstration of Association of Mouse Chromatin Accessibility with Heritable Human Traits from an Initial GWAS from the UK Biobank. A) Heatmaps show matrices of -log10(q-value) of enrichment. Trait/cluster pairs with no significant enrichment are shown in white. The proportion of cell that originate from each tissue is plotted alongside the heatmap. B) Selected sections from the above heatmap with an additional graph of cell counts for each cluster alongside the heatmap. C–F) Individual traits (columns) from the lower heatmap in panel B. Within each panel the following are shown for each cluster: the proportion of each tissue composing that cluster and the -log10(q-value) of enrichment. Clusters are sorted by q-value and the dotted line indicates a q-value of 0.05. Note that all traits additionally report an ID in parenthesis which corresponds to the ID given to each phenotype by the Neale Lab.

4.4 Discussion

Our study is limited in at least three ways. First, at the level of single cells, the data are very sparse, and generating "pseudo-bulk" profiles for each cell type requires aggregating data from many similar cells. At this stage, we are only powered to characterize relatively common cell types. Additional work (e.g., further scaling, protocol improvements, development of an RNA/ATAC co-assay) will be necessary to effectively profile chromatin accessibility in rare cell types.

Second, although we are reasonably confident in our cell-type assignments, they should be regarded as preliminary, especially as we show that such assignments can vary a great deal between independent studies. We were mostly in agreement with two recent scRNA-seq atlases, but the discrepancies are illuminating. For example, in our study, two clusters we labeled as podocytes and endothelial cells in the kidney and lung, respectively; both match "stromal cells" of Han et al. (2018) Similarly, a cluster we annotate as collecting duct in kidney matches "epithelial cells" of Han et al. (2018). While it is possible that our label-transfer method spuriously suggested these pairings, it seems more likely that independent annotation simply led to different conclusions. We anticipate that true errors, as well as discrepancies secondary to different "label resolution"

choices (e.g., collecting duct cells are epithelial), will occur in all atlas comparisons. The further development of robust methods to compare annotations across atlases and address discrepancies should be a high priority for the field.

Third, an immediate challenge that any molecular atlas of disaggregated cells faces is the choice of a cell/nucleus isolation protocol. Our cellular isolation protocol was generally robust, but not across all tissues (e.g., pancreas, skeletal muscle). As cells are routinely isolated from these tissues in other contexts, we expect that small optimizations will resolve this issue. In addition, for several tissues, even though we generated datasets that would pass typical QC, we are less confident in the data based on downstream analyses. In particular, sci-ATAC-seq of intestinal samples did not yield the expected biological variation besides one cluster of likely enterocytes.

We are excited about future possibilities on three fronts. First, we view this study as one of several that are laying the foundation for a comprehensive single cell molecular atlas of mouse. A major advantage of generating a mouse atlas, relative to a human atlas, is the possibility of accessing early developmental time points. Future studies will expand the number of tissues and time-points analyzed, the modalities of molecular information collected, and ideally will integrate with both lineage and spatial information Frieda et al. (2017); McKenna et al. (2016).

Second, there are many aspects of our data that remain incompletely explored. Future investigations might include (1) identifying the cell types of clusters that we failed to assign, (2) understanding how subtle differences in TF regulatory grammar underlie differential accessibility of specific elements between closely related cell types, (3) applying pseudo-ordering approaches to comprehensively identify correlates of heterogeneity (e.g., space, differentiation, cell state), or (4) building predictive models of cell-type-specific gene expression that are based on chromatin accessibility of linked distal elements. To these and other ends, we hope that our data will be broadly useful to the community (available at `http://atlas.gs.washington.edu/mouse-atac` along with vignettes to facilitate their exploration).

Finally, as illustrated by our analyses of mouse cell-type enrichment in heritability for diverse human traits, the generation of an equivalent single-cell atlas of chromatin accessibility from human tissues is well motivated. Furthermore, as was the case with the mouse and human genomes,

the lens of evolutionary comparison will be essential for maximizing the value of data from either species. As similar single-cell atlases of chromatin accessibility are generated and made available for humans and other species, we anticipate rich opportunities for investigating the evolution of cell-type-specific cis-acting regulatory elements, the evolution (or death thereof) of the trans-acting regulatory milieu within each cell type, and the birth and death of individual cell types (Arendt et al., 2016).

4.5 METHODS

All methods are published in Cusanovich et al. (2018). The majority of these methods are replicated below.

4.5.1 Tissue Extractions and Nuclei Isolation

In total, tissues were extracted from 5 mice across three different days. We successfully isolated nuclei from 13 different tissues: bone marrow, cerebellum, large intestine, heart, kidney, liver, lung, prefrontal cortex, small intestine, spleen, testes, thymus, and whole brain. The details for generating single cell suspensions from each tissue are as follows:

Bone marrow: both femurs were cut open and a 23 gauge needle was inserted into the small excised piece of bone to flush out bone marrow cells with 1 mL cold DMEM supplemented with 5% FBS.

Cerebellum, kidney, liver, lung, prefrontal cortex, testes, thymus, and whole brain: whole tissues were dissected and rinsed thoroughly in PBS before placing in 1 mL of cold DMEM supplemented with 5% FBS to await further processing on ice. Each tissue was then minced with a razor blade and then transferred to a 7 mL Dounce homogenizer to homogenize (with a loose and then tight pestle if needed) x 8-10 strokes in 5 mL of total volume cold DMEM with 5% FBS on ice.

Heart and Large Intestine (including cecum and colon): whole tissues were dissected and rinsed thoroughly in PBS before placing in 1 mL cold DMEM supplemented with 5% FBS to await further processing on ice. 3-4 mL of Accutase (Millipore SF006) was added and the tissue was minced

with a razor blade before incubation with rotation at 37C for 30 minutes followed by incubation at room temperature for 30 minutes. After digestion, 3 mL of cold DMEM with 5% FBS were added to each tissue and samples were homogenized using a 7 mL Dounce homogenizer (loose and then tight pestle if needed) x 5-8 strokes on ice to obtain a homogeneous cell suspension.

Small intestine (including duodenum, jejunum, and ileum): small intestine was dissected and rinsed thoroughly in PBS before placing in 1 mL cold DMEM supplemented with 5% FBS to await further processing on ice. 3-4 mL of Collagenase type 2 (Worthington Type 2, 1 mg/ml in Alpha MEM) was added and the tissue was minced with a razor blade before incubation with rotation at 37C for 2 hours. After digestion, 3 mL of cold DMEM with 5% FBS was added to digest and samples were homogenized using a 7 mL Dounce homogenizer (loose and then tight pestle if needed) x 5-8 strokes on ice to obtain a homogeneous cell suspension.

Spleen: after dissecting out the spleen, a 23 gauge needle was used to puncture the splenic sac and splenocytes were flushed out with 1 mL of cold DMEM with 5% FBS.

We also note that we tried to extract single cell suspensions from several other tissues (including skeletal muscle and pancreas), but were unsuccessful with these conditions. After isolating single cell suspensions for all of the samples, cells were pelleted by spinning at 500xg for 5 minutes at 4C. We then discarded the supernatant and washed the cell pellets by re-suspending in 10 mL cold PBS followed by centrifugation again at 500xg for 5 minutes at 4C. The cell pellets were then re-suspended in 1 mL cold lysis buffer (10 mM Tris-HCl, ph 7.4, 10 mM NaCl, 3 mM MgCl2, 0.1% IGEPAL CA-630 - from (Buenrostro et al., 2013) supplemented with protease inhibitors (Sigma P8340). If needed, samples were further homogenized on ice using a 2 mL dounce homogenizer to obtain a final homogeneous nuclei suspension. Nuclei were then incubated cold lysis buffer with protease inhibitors at 4C for 1 hour. Subsequently, nuclei were strained using 70 micron cell strainers to obtain a single nuclei suspension by placing a cell strainer onto the top of a 50 mL conical tube and pipetting the nuclei suspension into the strainer. The plunger of a 1 mL syringe was used to gently tap the strainer until the material was passed through. The strainer was then rinsed with an additional 0.5 mL of cold lysis buffer supplemented with protease inhibitors if necessary. Isolated nuclei were then stained with 3 μM DAPI and then DAPI+ nuclei were sorted

into 96-well plates (2,500 nuclei per well) containing 20ul of nuclear freezing buffer in each well (50 mM Tris at pH 8.0, 25% glycerol, 5 mM Mg(OAc)2, 0.1 mM EDTA, 5 mM DTT, 1X protease inhibitor cocktail [Sigma]) using a BD FACS Aria II and then flash frozen and stored at -80C. Each tissue was sorted onto a different set of plates.

4.5.2 Generating sci-ATAC-seq Libraries

To generate sci-ATAC-seq libraries, we followed a protocol similar to previously described experiments (Cusanovich et al., 2015, 2017), with a few modifications. At least 2 tissues were processed at a time, ensuring that replicates of the same tissue were not processed on the same date. Plates of frozen nuclei in nuclear freezing buffer were thawed and then 20 μl of TD buffer (Illumina, part of FC-121-1031) was added to each well. 1 μl of each of the 96 custom and uniquely indexed Tn5 transposomes (Illumina, 2.5 μM) (Amini et al., 2014) was then added to each well and nuclei were incubated at 55C for 30 minutes. Following tagmentation, 40 μl of 40 mM EDTA (supplemented with 1 mM Spermidine) was added to stop the reaction and the plate was incubated at 37C for 15 minutes. All wells of the plate were then pooled, nuclei were stained again with 3 μM DAPI and 25 DAPI+ nuclei were sorted into each well of a second set of 96-well plates that contained 12.5 μl of nuclear lysis buffer (11.5 μl of EB buffer [QIAGEN] supplemented with 0.5 μl of 100X BSA [NEB] and 0.5 μl of 1% SDS). For each tissue, up to 4 96-well plates of nuclei were collected for further processing. We have previously estimated that sorting 25 nuclei into each well of the PCR plates will result in approximately 12% of barcodes representing more than one nucleus ('collisions', (Cusanovich et al., 2015)). After sorting, samples were frozen at -20C until ready for PCR amplification. For amplification, we thawed plates and then added the first indexed PCR primer to each well (0.5 μM final concentration), incubated the samples at 55C for 15 minutes, and then transferred each well to a 384-well plate using the Liquidator 96 Manual Pipetting System (Mettler Toledo). Finally, we added the second indexed PCR primer (0.5 μM final concentration) and 7.5 μl of NPM polymerase master mix (Illumina, FC-121-1012) to each well. Tagmented DNA was then PCR amplified. To determine the number of cycles required, we first amplified several

test wells of nuclei that had been sorted onto an additional plate and monitored the reactions with SYBR green on a qPCR machine to establish when the libraries reached saturation. The cycling conditions were as follows: 72C 3 minutes, 98C 30 s, followed by 15-25 cycles of 98C 10 s, 63C 30 s, 72C 1 minute, and finally Hold at 10C.

The optimal number of cycles can vary from one experiment to the next. Based on our qPCR results, we amplified libraries for 19-22 cycles (depending on the tissue). After PCR amplification, all wells of each 384-well plate were pooled and cleaned up using a DNA Clean and Concentrator-100 column (Zymo) and then cleaned again using 1X Ampure beads (Agencourt). All steps that required pipetting nuclei were done with wide-bore tips. Finally, the concentration and quality of the libraries was determined using the BioAnalyzer 7500 DNA kit (Agilent). For sequencing, equimolar libraries from each of the 384-well plates were pooled and sequenced on two runs of a HiSeq 2500 (Illumina) and using custom primers and a custom sequencing recipe (Amini et al., 2014). 50 base pairs (bp) were sequenced from each end, in addition to the barcodes introduced during tagmentation and PCR amplification.

4.5.3 Raw Processing of Data

After sequencing, BCL files were converted to fastq files using bcl2fastq v2.16 (Illumina). Each read was associated with a cell barcode made up of 4 components: on the P5 end of the molecule there was a row address for tagmentation and for PCR added and on the P7 end of the molecule there was a column address for tagmentation and PCR added. To correct for errors in these barcodes, we broke them into their 4 constituent parts and calculated the edit distance for each piece from all possible barcode addresses. If an individual component was within 3 edits of an expected barcode and the next best matching barcode was at least 2 edits further away, we corrected the barcode to its best match. Otherwise, the barcode component was classified as ambiguous or unknown.

We next trimmed reads with Trimmomatic (Bolger et al., 2014) and then mapped reads to the mm9 reference genome using bowtie2 with $-X2000-31$ as options (Langmead and Salzberg,

2012) and then filtered out read pairs that did not map uniquely to autosomes or sex chromosomes with a mapping quality of at least 10 using Samtools (Li et al., 2009), as well as reads that were associated with ambiguous or unknown barcodes. Of 3,895,812,907 sequenced read pairs, 2,598,089,519 (67%) mapped to the nuclear reference genome, with an assigned cell barcode. In contrast, only 14,341,775 read pairs (0.4%) mapped to the mitochondrial genome, with an assigned cell barcode.

We subsequently removed PCR duplicates for all reads that mapped to the nuclear genome using a custom python script that removes duplicates on a cell-by-cell basis. We then separated out the reads from each tissue for further processing. Next, in order to identify barcodes representing genuine cells we counted the number of reads assigned to each barcode, which is bimodally distributed when log-transformed representing a low read depth distribution of background reads assigned to improper barcodes and a high read depth distribution of reads deriving from real cells, and used the mclust package (Scrucca et al., 2016) in R to distinguish between the two populations of barcodes. We performed this procedure for each tissue separately, setting the read depth cutoff for a cell at the point at which there was no more than 5% uncertainty that the barcode belonged to the high read depth distribution (Figure 4.1D).

One hallmark of a successful bulk ATAC-seq library is a periodicity in the frequency of insert sizes for the sequenced molecules, reflecting the fact that Tn5 does not insert into nucleosome associated DNA as efficiently. We noted that even with very few reads, we could observe this periodicity in individual cells (Figure 4.1D), and so we decided to filter out individual cells with poor signals of nucleosomal periodicity in their fragment size distribution. To do so, we calculated a periodogram using a fast Fourier transform of insert sizes for each cell using the spec.pgram() function in R with $pad = 0.3, tap = 0.5, span = 2$ and then summed the spectral densities for frequencies between 100 and 300 bp as a measure of the strength of the nucleosomal signal. We found that the log-transformation of these scores across all tissues was roughly normally distributed with a long lower tail (representing cells with increasingly poor nucleosome signals) and so we only retained cells with a log-transformed score of -0.8 for further analysis.

4.5.4 Latent Semantic Indexing Cluster Analysis

In order to get an initial sense of the relationship between individual cells, we first broke the genome into 5kb windows and then scored each cell for any insertions in these windows, generating a large, sparse, binary matrix of 5kb windows by cells for each tissue. Based on this matrix, we retained the top 20,000 most commonly used sites in each tissue (this number could extend a little above 20,000 because we included tied sites at the threshold) and then filtered out the bottom 5% of cells in terms of the number of 5kb windows with any insertions. We then reduced the dimensionality of these large binary matrices using a term frequency-inverse document frequency ("TF-IDF") transformation. To do this, we first weighted all the sites for individual cells by the total number of sites accessible in that cell ("term frequency"). We then multiplied these weighted values by log(1 + the inverse frequency of each site across all cells), the "inverse document frequency." We then used singular value decomposition on the TF-IDF matrix to generate a lower dimensional representation of the data by only retaining the 2nd through 10th dimensions (because the first dimension tends to be highly correlated with read depth). These LSI-transformed scores of accessibility were then standardized by row (i.e., mean subtracted and divided by standard deviation), capped at $\pm$ 1.5, and used to bi-cluster cells and windows based on cosine distances using the ward algorithm in R. Visual examination of the resulting heatmaps identified between 2 and 7 distinct clusters of cells, depending on the tissue. These relatively crude groups of cells were used for peak calling (described below) to maintain enough cells in each group for identifying peaks while also retaining sufficient sensitivity to identify peaks that were restricted to subset of cells.

4.5.5 Identifying Peaks of Accessibility

On the basis of the crude clustering of cells above we next identified specific regulatory elements that were accessible in each cluster. To do so, we combined the data across cells from each cluster to generate aggregated profiles of accessibility for groups of cells (a process we refer to as "in silico cell sorting"). For this analysis we simply collected all the unique mapped reads associated with cells that were assigned to a given cluster within a tissue and saved that as a distinct bam file.

Then for each bam file representing a cluster, we used MACS2 (Zhang et al., 2008) to identify peaks of increased insertion frequency. Specifically, we used the macs2 callpeak command with the following parameters: $--nomodel --keep -dupall --extsize200 --shift -100$. For downstream analyses we generated a master list of potential regulatory elements by taking all peaks identified in any cluster in any tissue sample and merged them with the BEDTools program (Quinlan and Hall, 2010).

For Figure 4.2, we also compared our sci-ATAC-seq data to DNase-seq bulk data collected by the ENCODE consortium where we had profiled the same tissue. In order to be consistent in our comparisons, we downloaded the raw DNase-seq fastq files (36 bp, single-end, `https://www.encodeproject.org/`) for all replicates from the same tissues we had profiled, remapped them with our pipeline and called peaks with MACS2 as described above. Specifically, we downloaded data for cerebellum (1 replicate), heart (1 replicate), small intestine (2 replicates), kidney (2 replicates), liver (14 replicates), lung (3 replicates), spleen (2 replicates), thymus (2 replicates), and whole brain (7 replicates). To facilitate quantitative comparisons across all of these tissues, we merged all peaks identified in our sci-ATAC-seq samples (across each entire tissue) and the ENCODE samples with BEDTools and then used the deepTools program (Ramírez et al., 2016) to count reads overlapping each peak for each tissue. These tissue-level peak calls were used to generate Figure 4.2A and to quantify the fraction of reads in peaks ('FRiP') reported in Figure 4.2B. To calculate FRiP scores, we divided the number of reads for each tissue that overlapped the union of all peaks identified in the DNase-seq and ATAC-seq libraries (overlap determined by BEDTools) by the total number of nuclear genome mapped reads. To generate the heatmap in Figure 4.2J we used the inverse of the spearman correlation between pairs of samples as a distance metric and ward clustering to cluster samples.

4.5.6 t-distributed Stochastic Neighbor Embedding and Iterative Cluster Analysis

To take a more holistic approach to understanding the relationships of different cell types across the entire dataset, we combined all cells from all tissues and used the t-distributed stochastic neighbor

embedding dimensionality reduction technique to visualize the full dataset and identify clusters of cells representing individual cell types. As with the LSI analysis above, we started by generating a large binary matrix of sites by cells, but instead of scoring cells for reads overlapping 5kb windows in the genome we scored cells for reads overlapping the master list of potential regulatory elements we had previously identified based on LSI clusters. Starting with all cells that passed our nucleosome signal and read depth thresholds, we again wanted to remove the most sparsely sampled sites and cells to more clearly define differences between cell types. To do so, we first filtered out any sites that were not observed as accessible in at least 5% of cells in at least one LSI cluster and then filtered out cells that were more than 1 standard deviation below the mean number of sites observed. We then transformed this matrix with the TF-IDF algorithm described above. Finally, we generated a lower dimensional representation of the data by including the first 50 dimensions of the singular value decomposition of this TF-IDF-transformed matrix. This representation was then used as input for the Rtsne package in R (Krijthe, 2015). To identify clusters of cells in this two dimensional representation of the data, we used the Louvain clustering algorithm implemented in Seurat (Satija et al., 2015). Resolution and K parameters for Louvain clustering were chosen for each major cluster to produce reasonable groupings of cells that are well-separated in each t-SNE embedding. This analysis identified 30 distinct clusters of cells, but to get at even finer structure, we subset TF-IDF normalized data on each of these 30 clusters of cells and repeated SVD and t-SNE to identify subclusters, again using Louvain clustering. Through this round of "iterative" t-SNE, we identified a total of 85 distinct clusters. Note that for one major cluster, major cluster 12, we found that Monocle 2's implementation of density peak clustering (Qiu et al., 2017b; Trapnell et al., 2014) seemed to produce more reasonable clusters. Rho and delta parameters were set in the same manner as for Louvain clustering.

4.5.7 Identifying Differentially Accessible Sites and Calculating Cell Type Specificity Score

To identify regulatory elements that were accessible in individual cell types, we used a logistic regression framework to test whether cells of a given clade were more likely to have insertions at

a given site relative to a reference panel of cells sampled uniformly from each tissue. To generate this reference panel we randomly sampled 120 cells from each of the 17 tissue samples collected in this study. We then used the differential test implemented in the Monocle 2 package (Qiu et al., 2017a; Trapnell et al., 2014) using the "binomialff" test with the following model:

$$logit(p_{ij}) = \mu_i + \alpha_j + \beta_j + \varepsilon_i$$

Where p is the probability that the ith site is accessible in the jth cell, μ is the total proportion of cells that are accessible at the ith site, α indicates the membership of the jth cell in the cluster being tested, β is the log10(total number of sites observed as accessible for the jth cell), and ε is an error term for the ith site. We used a likelihood ratio test framework (as implemented in Monocle 2) to determine if the full model (including cell cluster membership) provided a significantly better fit of the data than a model that only accounts for the intercept and the log10(number of sites observed in each cell). We set a 1% FDR threshold (Benjamini-Hochberg method) to determine whether sites were significantly differentially accessible for each of the 85 cell clusters relative to the reference panel of cells. For this analysis, only sites observed in at least one cell in a given cluster were tested.

Taking this a step further, we also defined sites with patterns of accessibility that were restricted to a limited number of cell types. For this, we used a specificity score that relies on Jensen-Shannon divergence similar to one that has been implemented in Monocle (Cabili et al., 2011; Trapnell et al., 2014). First we calculated the proportion of cells of each cluster that were accessible at each site. To make these proportions more comparable across clusters, which may be represented by cells with different overall complexity, we calculated a scaling factor for each cluster. To do so, we first calculated the median number of sites accessible in individual cells for each cluster. After log10-transforming these values took the ratio of the average median accessibility (across all clusters) over the median accessibility in each cluster as our scaling factor. The proportions in each cluster were then multiplied by these factors to arrive at cluster complexity-corrected proportions. We then re-scaled these normalized proportions on a site-by-site basis so that they were a probability distribution, representing the fraction of the maximum normalized proportion observed in any

cluster for each site:

$$p1 = \frac{\text{normalized proportions}_{\text{site i}}}{\sum_{i=1}^{n}(\text{normalized proportions}_{\text{site i}})}$$

we then calculate the Jensen Shannon divergence between two probability distributions

$$\text{JS divergence } = \frac{H(p1+p2)}{2} - \frac{H(p1)+H(p2)}{2}$$

where

$$H = -\sum_{i=1}^{n} p_i * \log_2(p_i)$$

and

$$p2 = \begin{cases} 1, \text{ current cluster} \\ 0, \text{ otherwise} \end{cases}$$

and finally we calculate our Jensen-Shannon-based specificity score as follows:

$$\text{JS specificity } = 1 - \sqrt{\text{JS divergence}}$$

Using this method we tested all 85 clusters for specificity (i.e., restricted patterns of accessibility) for all 436,206 master regulatory elements. To ensure that we identified sites that are commonly accessible within a given cluster of cells and rare in other clusters and not just sites that are rarely accessible overall with this analysis, we further transformed these Jensen-Shannon specificity measures by squaring them and then multiplying them by the normalized proportion of cells a site is accessible in for a given cluster of cells. In order to determine a reasonable cutoff for these specificity scores, distinguishing restricted accessibility patterns and more global patterns, we employed an empirical FDR-like strategy. To do this, we considered all site/cluster tests that were not significantly differentially accessible (from the likelihood ratio tests above) to be a null distribution of specificity scores and all specificity scores for site/cluster tests that passed our likelihood ratio test threshold and had positive beta values to be true positives. We then set the threshold for specificity such that no more than 10% of the site/cluster tests with larger specificity scores than

this threshold came from the null distribution. Finally, we filtered out the 10% of sites/cluster pairs from the null distribution that passed this specificity threshold (i.e., we required that a site/cluster pair had to pass the specificity score threshold and the differential accessibility threshold).

4.5.8 Linking distal sites to putative target genes

We defined a site as distal if it was located >5 kb upstream and >1 kb downstream of any transcription start site (TSS) reported in GENCODE mm9, release M1. We only considered sites that were accessible in at least 1% of the cells in each cluster as open in that cluster. We then ran Cicero on open sites for each cluster separately to identify co-accessible sites using the following parameters: aggregation k = 30, window size = 500 kb, distance constraint = 250 kb. The aggregation value k is the number of cells that are aggregated using k-nearest neighbors prior to calculating co-accessibility scores, the window size parameter controls the size of each model window in the genome, and the distance constraint parameter is the distance at which the distance penalty is trained to regularize the majority of connections. Using Cicero, we were able to find sites that were co-accessible in aggregated groups of cells within each cluster. Cicero assigns a regularized co-accessibility score to each pair of "open" sites in each cluster using a Graphical Lasso model, penalized by genomic distance (Pliner et al., 2018).

We first used Cicero maps to find a global *cis*-regulatory view of genome in different clusters with a co-accessibility cutoff of 0.2. Using this threshold and the windows around any TSS defined above, we were able to define pairs of sites that were co-accessible into proximal-to-proximal, distal-to-distal, and distal-to-proximal linked sites. Second, we used these co-accessibility maps to perform enrichment tests to help inform our biological interpretations of each cluster. For this purpose, we focused on distal differentially accessible (DA) sites and proximal DA sites in each cluster. In order to assign DA distal sites to target genes, we devised the following linking policy: i) distal DA sites were associated to any proximal site if they had a co-accessibility cutoff >0.2; ii) if a distal DA site was not linked to any proximal site with a co-accessibility cutoff of 0.2, it was assigned to the proximal site with highest co-accessibility, provided that this co-accessibility score

was greater than a relaxed cutoff of 0.1. The union of genes linked to distal DA sites under the relaxed policy and proximal DA sites were used for annotation enrichment tests.

We used Piano package (Väremo et al., 2013) for performing gene annotation enrichment analyses on following ontologies downloaded from `http://download.baderlab.org/EM_Genesets/June_20_2014/Mouse/symbol/`: all pathways: Mouse_AllPathways_June_20_2014_symbol.gmt GO biological process: MOUSE_GO_bp_with_GO_iea_symbol.gmt REACTOME: Mouse_Reactome_June_20_201

We also curated mouse a phenotype gene set from the MGI phenotype repository MGI_PhenoGenoMP.rpt (Bello and Eppig, 2016). The corresponding gmt file is available on our website. The "runGSAhyper" function in the Piano package was used to perform hypergeometric tests on the annotation terms associated with genes linked to distal DA sites and/or proximal DA sites in each cluster. The background for these tests was set to all genes with a proximal peak in this study. Terms that did not have at least 5 gene hits in the test set were filtered out. In addition, we defined a term significant if it had fold change (observed/expected) ¿= 2 and q-value ¡ 0.0001. The resulting terms for all the clusters were aggregated and were visualized as heatmaps (see Additional Resources).

4.5.9 Computing gene activity scores

For each gene in each cell, we compute "activity scores" which summarize the degree to which chromatin surrounding the gene is accessible. We do so using Cicero, which includes a procedure to summarize overall chromatin accessibility of all sites it links to a given gene (Pliner et al., 2018). Details are reproduced here for clarity. Briefly, the method works as follows: first, Cicero calculates an overall measure of the accessibility of sites linked to each gene k by first selecting rows of the binary accessibility matrix A that correspond to sites proximal to the gene's transcription start sites or to distal sites linked to them. These rows are weighted by their co-accessibility and then summed to produce a vector of accessibility scores R_k, where the overall accessibility of gene k in cell i is:

$$R_{ki} = \sum_{p \in P} \sum_{j \in P_p} A_{ji} \frac{u_{pj}}{\sum_{k \in D_p} u_{pk}} + A_{pi}$$

where P indexes the promoter proximal sites of k, D_p indexes distal sites linked to proximal site p, and u is the Cicero co-accessibility score linking distal site j to proximal site p, and A is the binary score for accessibility at site j or p in cell i.

Because the magnitude of these aggregate accessibility values will depend on overall sci-ATAC-seq library depth in each cell, we capture this relationship via a linear regression:

$$\log\left(\sum_k R_k\right) = \beta_0 + \beta_A \log\left(\sum_j A_{ji}\right)$$

The aggregate accessibility for each gene k in cell i is then scaled using the output of this model r_i for cell i:

$$\tilde{R}_{ki} = R_{ki} \cdot \frac{\sum_i r_i}{r_i}$$

Gene expression values measured by RNA-seq are typically approximately log-normally distributed. We therefore transform aggregate accessibility values to gene "activity" scores G_{ki} for each gene k in each cell i by simply exponentiating them. We also scale them by the total (exponentiated) gene accessibility values to produce "relative" activities:

$$C_{ki} = \frac{e^{\tilde{R}_{ki}}}{\sum_k e^{\tilde{R}_{ki}}}$$

The distribution of activity values of all genes in a given cell typically resembles a bimodal distribution, the lower peak of which corresponds to genes that are very lowly or not expressed in the population of cells. In contrast, genes in the second peak are typically expressed at appreciable levels in the population of cells. To ensure that non-expressed genes receive an activity score of zero, we first compute the mean activities for each gene across all cells in each cluster, and then fit a mixture of two Gaussians to the mean values. The larger of the location parameters is used as a threshold; all activity values in all cells in the cluster are divided by the threshold and rounded to

the nearest integer. Finally, we normalize these activity values by computing "size factors" using Monocle 2 with previously described methods for normalizing scRNA-seq data (Qiu et al., 2017b).

4.5.10 Classifying Clusters by Cell Type

To identify the cell type or types found in each chromatin accessibility cluster, we first collated a set of cell types expected in each tissue (Ross and Pawlina, 1979). Next, we compiled a list of marker genes for each cell type that are either commonly used markers (e.g., in immunohistochemistry experiments) or are crucial for carrying out functions believed to be exclusively performed by that cell type.

We next assessed the specific activity of our marker panel as follows. We devised a test for differential gene activity by applying the regression framework used to test sites for differential accessibility to the gene activity scores described in the Identifying Differentially Accessible Sites and Calculating Cell Type Specificity Score section. Doing so required that we binarize the activity scores, which we did simply by treating all activity values $C_{ki} > 0$ as "active" and all zero values as inactive. This test yielded the set of genes g for each cluster j for which being a member of j significantly increased the log-odds (denoted α) it was active (FDR ¡ 1%). We then intersected the set of genes g with our curated set of markers. To identify likely cell types found in a cluster j, we ranked the significantly predictive genes by their test scores αk. This scheme ranks markers that are predictive only for cluster j higher than those that are predictive for j as well as other clusters. However, collisions, low-quality or low-depth libraries, and technical noise could result in a cluster comprised mainly of one cell type being contaminated with a small number of cells of another type. In this case, markers from the contaminating cell type might appear to be significantly predictive of the cluster as well. We found penalizing marker activity scores according to the proportion cells from each tissue to be effective means of preventing misclassifications from cluster misassignment. Specifically, for each marker, we compute a weight:

$$w_k = \sum_t p_t$$

where p_t is the proportion of cells in the cluster from tissue t that contains cell types for which k

is a marker. Having identified and ranked significantly active markers for each cluster, our pipeline suggested up to three cell types for each cluster (based on the top three most active markers), and provided them as part of an automated marker report (see Additional Resources).

To refine cases where the report above included assignments to more than one cell type, we used a classifier-based approach to attempt to identify a single likely cell type assignment. The report described above containing possible cell types for each of the 85 clusters along with a set of markers for each was used to identify cells from each cluster that had a non-zero value of binariazed gene activity for markers from exclusively one of the cell types listed for its cluster. These cells were used as training examples for a logistic regression model using scikit-learn using all genes as input (excluding the initial markers), an L1 penalty, a value of $C = 1$ for regularization, and $class\_weight = \text{"}balanced\text{"}$. This model was then used to classify all cells in the dataset and prediction probabilities were calculated using the function *predict_proba* and the model. Assignments were only considered valid if the predicted class had greater than five times higher probability than the next most probable class. Clusters of cells having 90% or more of the assignments of individual cells to a single class were given the majority assignment and all others were given an assignment of unknown.

Finally, we also manually reviewed each assignment and in a few instances modified automated cell type assignments. For the manual review of these assignments we considered an in-depth examination of markers, clustering of groups of cells in the differential accessibility heatmap in Figure 4.3D, and conclusions drawn from more in-depth analysis of specific subsets of cells (e.g., Figure 4.10, Figure 4.11, Figure 4.12). For some small clusters, we observed that that their accessibility profile appeared to be a combination of two other cell types and so we have labeled them as "collisions." In addition, several clusters did not have a clear assignment and have been labeled as "unknown." The results of the automated assignment and manual provided us with a final assignment for each cluster that was used throughout the results. A table of the automated assignments, our manual curation, and notes on the rationale of any manually modified cell type assignments (including notes on "collision" assignments) are provided in supplementary data files.

4.5.11 Obtaining External scRNA-seq Datasets

scRNA-seq data of adult mouse kidney from Park et al. (2018) were obtained from GEO: GSE107585. All cells included in this matrix were included in the analysis and cluster assignments from the provided matrix were mapped to the cell type assignments provided in the text of Park et al. (2018).

scRNA-seq data of several organs were obtained from Han et al. (2018), via FigShare `https://figshare.com/s/865e694ad06d5857db4b`) and the table containing cell type assignments for clusters and cluster assignments were merged and then associated with the scRNA-seq data with batch removed as reported by the authors. Note that we observed similar results for the analyses presented here when using versions with and without batch removal. The authors report cell type assignments for roughly 260K cells of the over 400K with expression data from their entire dataset, so only expression profiles corresponding to one of these cells were used in analysis. We further filtered to cells with at least 500 UMIs and less than 10% mitochondrial transcripts as described in Han et al. (2018).

Data from the Tabula Muris project were downloaded from FigShare (`https://figshare.com/projects/Tabula_Muris_Transcriptomic_characterization_of_20_organs_and_tissues_from_Mus_musculus_at_single_cell_resolution/27733`). This final version used for figures was the V2 version of this dataset uploaded on 03/27/2018. Assignments as noted in annotation files were associated with scRNA-seq data and used as reported on FigShare. Only cells with assignments were used in analysis. Only datasets from 10X chromium were used, not FACS sorted well-plate samples.

4.5.12 Preprocessing of Cell Type Labels from Han et al. 2018

For the purposes of downstream analysis, we opted to condense some of the labels provided in Han et al. (2018). We made this decision because in inspecting the data, there were many groups that received the same cell type assignment as one another but the names reflected notation of the gene that was most highly differentially expressed within that particular group. We were most interested in comparing groups that reflected readily distinct cell types, so we removed these minor

designations from the labels prior to analysis and other groups of cells that appeared to be very similar in name and expression profile were also combined into a single category in an effort to make comparisons of more appropriately matched resolution across each of the three scRNA-seq datasets noted above.

For the two remaining scRNA-seq datasets, the labels were used as provided by the authors without modification.

4.5.13 Comparison of Activity Scores and Assignments to scRNA-seq Data

To compare our activity scores to relevant published scRNA-seq datasets mentioned above, we used two different methods. The first is a simple correlation-based approach. To do so, we subset to the set of genes common to both datasets. We also filtered to a set of labels with a frequency of 0.5% or more within either dataset. To select features, we first we created separate Seurat objects (using the Seurat R package) using the expression / activity score values for those cells, analyzing the ATAC and RNA datasets separately, and then normalized the data using the NormalizeData function in Seurat. We then estimated the top 3000 most variable genes in the ATAC and RNA datasets separately using the function FindVariableGenes. We then combined the ATAC and RNA data (using the full set of intersecting genes, not restricting to variable genes) into a single Seurat object and normalized the combined data as described above and scaled using the ScaleData function in Seurat. We calculated average normalized expression / activity score profiles (as normalized by Seurat) for each annotated group of cells within each dataset. We then calculated Spearman correlation coefficients between average profiles of all pairs of groups as a metric of similarity, restricting to the top 3,000 variable genes derived from the scRNA-seq dataset. This was done to ensure that we measure concordance with what one would typically see in existing scRNA-seq datasets.

In an attempt to develop a method that could be used to annotate individual cells independent of clustering, we performed the same set of initial steps, but rather than calculating average normalized expression profiles, we performed PCA where each PC is weighted by variance explained (50

PCs calculated) using the function RunPCA in Seurat. Note that for this step, we restricted to the top 3,000 most variable genes as measured by sci-ATAC-seq via activity scores as calculated above. We find that this performs moderately better than using variable genes as defined by scRNA-seq, likely due to some genes whose variation is still not captured sufficiently by activity scores. Within this PCA space we calculate the 20 nearest neighbors of each cell in the ATAC dataset within the RNA dataset. We note that performing KNN within this PCA space rather than the raw space substantially speeds computation and generally provides improved performance in our experience. The most common majority label from the RNA neighbors (¿ 6 cells) is then assigned as the label for each ATAC cell (or NA if no label meeting this threshold is found).

Note that for the KNN strategy, we restricted to only ATAC cells with at least 1,800 sites measured per cell and RNA cells with at least 600 UMIs measured in set of genes common to both datasets. We found that this improved performance, while retaining the majority of cells. We found this to be particularly relevant when comparing to the datasets from Han et al. (2018), where the number of UMIs per cell is relatively low on average. The cutoff used for the scRNA-seq data here is still quite low compared typical complexity for existing large-scale single cell RNA-seq datasets, so we expect this choice of threshold is quite reasonable.

We also note that of the options we tried, activity scores seemed to work the best, but that similar comparisons using correlations or aggregate measures of reads or reads in peaks surrounding the promoter we also promising and may represent a simpler alternative to be explored further.

4.5.14 Trajectory analysis of hematopoiesis

To further investigate the chromatin accessibility landscape of hematopoiesis, we took the subset of cell clusters from Cusanovich et al. (2018) where a high percentage of cells were derived from the bone marrow (10, 13, 17.4, 24 and 28) for further analysis. We set out first to order cells in a branched trajectory, as we expected blood progenitors at various stages to be present in the marrow. To order cells, we used a modification of the Monocle 2 pseudotime ordering algorithm, as was used in Pliner et al. (2018). Briefly, peaks of accessibility within 10 kb of each other were

aggregated from the binary matrix to create a count matrix. Cells were clustered using density peak clustering after tSNE dimensionality reduction, and then a differential accessibility test was performed for each site, including the cluster labels as indicator variables (full model of cluster and reduced model of 1), to find differentially accessible sites across clusters. The top 3,000 differentially accessible sites by adjusted p value were chosen as ordering genes for trajectory inference. We then used Monocle 2's DDRTree dimensionality reduction algorithm to align cells to a branched trajectory as shown in Figure 4.12.

To validate the trajectory inferred using Monocle 2, we compared our tree with H3K4me1 ChIP-seq data on sorted hematopoietic cell populations from **?**. To do this, we summed a TF-IDF normalized matrix of read counts from sci-ATAC-seq peaks that overlapped H3K4me1 ChIP-seq from each sorted cell population. The resulting 'Enhancer accessibility' from each cell population was plotted in Figure 4.12B. The myeloid, erythroid, and lymphoid accessibilities resulted from the sum of scores from these populations and their progenitors.

We next wanted to examine the well-studied β-globin locus at various points along the erythroid lineage. Cells were divided into 5 time points with equal numbers of cells based on the assigned pseudotime and co-accessibility was calculated on each set of cells separately using Cicero (Pliner et al., 2018).

4.5.15 Calculating Enrichment of Heritability Measured by GWAS within Cell-type Peaks

To estimate enrichments of heritability for various complex traits in differentially accessible peaks for each cluster of cells we used LDSC, which takes summary statistics from a given GWAS as input and quantifies the enrichment of heritability in an annotated set of SNPs conditioned on a baseline model that accounts for the non-random distribution of heritability across the genome. To integrate human and mouse data, we first used the UCSC utility liftOver to lift all GWAS SNPs that are used by the LDSC software (`https://github.com/bulik/ldsc`) to the mouse genome (this is done when making the template files that are used as input to partitioned LDSC). We then took the set of differentially accessible peaks (in the positive direction) for each cluster

and annotated each SNP according to whether or not it overlapped one of these peaks. We then followed the recommended workflow for running LDSC using HapMap SNPs, precomputed files corresponding to 1000 genomes phase 3, excluding the MHC region to generate an LDSC model for each chromosome and peak set (see LDSC documentation).

To calculate enrichments based on each model, we first regenerated the baseline model (version 1.1) provided via the LDSC website and used this as the reference for enrichment calculation. We obtained full summary statistics for most GWAS used in Figure 7 from the Broad LD Hub (`https://data.broadinstitute.org/alkesgroup/sumstats_formatted/`), which also contains information on each individual study. Additional studies on asthma (GABRIEL; `https://beaune.cng.fr/gabriel/gabriel_results.zip`), chronic kidney disease (NHLBI; `https://fox-nhlbi-nih-gov.offcampus.lib.washington.edu/CKDGen/formatted_round3meta_CKD_overall_IV_2GC_b36_MAFget005_Nget50_20120725_b37.csv.gz`), IgA deficiency (`ftp://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/BronsonPG_27723758/BronsonEtAl_NatGenet_2016.zip`), and reproductive phenotypes (`ftp://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/BarbanN_27798627/`) were downloaded and processed separately with munge_sumstats.py provided with the LDSC software. The script ldsc.py from the LDSC github page was then used with default parameters and the baseline model above to calculate enrichments.

Results for all trait/cluster pairs were gathered into a single file and only traits with an estimated heritability of 0.01 or higher were carried forward for analysis. P-values were calculated from z-scores assigned to coefficients reported by ldsc.py and coefficients were divided by the average per-SNP heritability for trait associated with a given test (as calculated from the number of SNPs and overall heritability reported in the .log files from ldsc.py), as recommended by the LDSC authors, producing scaled coefficients. These scaled coefficients are provided in supplemental data files but are not reported in figures here. Tests were corrected for multiple hypothesis testing using the Benjamini-Hochberg method and only tests with a q-value of 0.05 or lower were deemed to be significant.

For the analysis performed on tissue peaks rather than cluster DA sites, peaks were called on

each tissue individually using the same peak calling strategy as outlined in "Identifying Peaks of Accessibility." These peaks were taken as the input set of peaks to the workflow described above.

4.5.16 UK Biobank Analysis

An initial set of GWAS results on UK biobank data was released by the Neale Lab, which we downloaded using the wget commands provided in the file: `https://www-dropbox-com.offcampus.lib.washington.edu/s/oe5q85454vhc3hi/phenosummary_final_11898_18597.tsv?dl=0`. More information about the GWAS performed and the files available for download are available on the Neale Lab website here: `http://www.nealelab.is/blog/2017/7/19/rapid-gwas-of-thousands-of-p`

After downloading these files, they were converted to the required input format for ldsc.py. The LDSC workflow was identical to the above with two main exceptions. First, when running ldsc.py we set the parameters–chisq-max and–two-step to 9999 (an arbitrarily high value) per recommendation of the Neale Lab. This is meant to adjust for the very large sample size of the UK Biobank. Second, we additionally excluded traits with an effective sample size of less than 5000, again, based on recommendations from the Neale Lab in the post above for obtaining stable LDSC results. As the Neale lab documents in their post linked above, for quantitative traits the effective sample size is simply the number of non-missing observations, while for case-control phenotypes this is calculated as effective_n = 4/((1/N.cases) + (1/N.controls)).

4.5.17 Convolution Neural Network Analysis

To identify sequence motifs enriched in different cell clusters we used Basset (Kelley et al., 2016), a convolutional neural network approach. We set the input to the CNN as the 600 bp sequences centered at the midpoint of the differentially accessible peaks for each cluster. The output of the classifier is a binary vector of length 85 (corresponding to the number of cell clusters). If a peak was significantly open in sub cluster i, the ith element of the output vector would be 1. the input sequence vectors were then converted to a matrix of 4 x 600 using a one-hot encoding strategy. We observed that the number of differentially accessible sites varied substantially across clusters,

with the median of 10,984. Some cell clusters had less than 1,000 DA sites and some had more than 30,000. This sort of imbalance can cause the CNN model to perform poorly for clusters with fewer DA sites, mainly because the CNN is not provided with enough sequences for those clusters during training. Therefore, we augmented the classes that had less than 10,984 DA sites by uniformly oversampling these sites. This forces all clusters to have a minimum 10,984 DA sites for training. For clusters with more than 10,984 DA sites, we only retained 10,984 for training by sorting DA sites by their beta values and then taking the top 10,984 with the largest effect sizes. We then randomly split the augmented dataset into 50% of sites for training, 25% for testing, and 25% for validation. We then used the Basset framework for training the CNNs. We used default Basset values for most parameters, except that we set the number of first layer filters to 600, each with width 19 and height 4, and we dropped the learning rate if the validation loss was not improving in 10 consecutive epochs. The trained model is available in Additional Resources. Although the main text only refers to this model, we also provide three additional trained models: (1) a model trained only on the clusters with ¿ 11,000 DA sites using all the DA sites in those clusters (imbalanced design), (2) a model trained only on clusters with ¿ 11,000 DA sites, but only considering the top 11,000 DA sites (ranked by beta value from the DA test) per cluster (balanced design), and (3) an unaugmented model trained on all the DA sites for each cluster, regardless of the number of DA sites identified. We note that evaluating the performance of these models is not straightforward, because this is an imbalanced classification problem where individual features can belong to many classes. This means that traditional measures of model performance will not be accurate - the area under the receiver operating characteristic curve ('AUROC') will tend to overestimate performance because of the imbalance, while the area under the precision recall curve ('AUPR') will tend to underestimate performance because features can belong to multiple classes. However, our aim here is to infer the sequence features that are most influential rather than to predict cluster membership directly. In other words, the drop in performance is interpretable even if the the overall performance of our trained model is poor (analogous to how likelihood ratio tests can identify differentially expressed genes even if the actual model fit is poor).

Next, we converted filter weights to PWMs by scaling the weights from 0 to 1 by dividing each

by the max weight in that filter. TomTom was used to annotate the filter PWMs with q-value cutoff of 0.1. TomTom identified known motif matches for 278 filters. We next evaluated the influence of the first layer filters on the trained CNN using a leave-one-out strategy. Among the 600 filters used in the first layer of the CNN, 20 had standard deviation of zero across the test dataset and so they were dropped from the analysis. For the remaining 580 filters, we calculated the influence of the filters on the prediction accuracy of each class, by dropping the filter and then calculating the loss using basset_motif_infl.py in Basset, where positive loss means that this filter was "influential" on this class. We then extracted the influence of known filters on each class. In order to calculate the influence of TF motifs on each class, we subset the influence table for filters with positive influence on any class. Because more than one filter might match the same motif, the influence of a motif on a class was set to max influence of filters matching that motif for that cluster. The resulting influence matrix of motifs on the corresponding 85 clusters were visualized as a heatmap in Figure 4.10B.

In order to extend the CNN results to single cells, we developed a simple procedure of matrix multiplication:

$$C_{ij} = A_k \times B_{kj}$$

where A is the binary matrix of cells by sites and B is a matrix of sites by motifs. To construct this site by motif matrix, we scanned sites for PWMs from the 580 filters of the first layer of the CNN that had non-zero standard deviation using FIMO (Grant et al., 2011) and constructed matrix B such that if site k had any match to filter j with a P-value ¡ 0.00001, we set Bkj as 1, otherwise 0. Using this method, we end up with a cell by motif matrix where the value for each cell/motif pair is an aggregated score for each cell of all sites that are accessible and contain a given motif. We then normalized these motif scores for each cell by cell size factor (defined as the median of ratios of read counts in a cell compared with all the cells).

Motif activity scores were rescaled by dividing to the max and then were projected on the t-SNE maps (Figure 4.10C). For visualizing the influence of individual motifs on chromatin profiles of individual cells, we chose the filter that best matched the motif of interest with a TomTom q-

value of at least 0.01 (multiple filters often matched to a known motif). These filters also matched to other motifs at a more relaxed q-value cutoff of 0.1. In Figure 4.10, filter 357 best matched to GATA motifs, but also matched to STAT1 and IRF4 motifs. Filter 361 best matched to the MEF2 motif but also matched to STAT5A and ARI3A motifs. Filter 36 best matched to the PPAR motif, but also matched to COT1, COT2, and NR4A3 motifs.

4.6 Data availability

The accession number for the sequencing data and some processed data files reported in this chapter is GEO: GSE111586.

We also provide extensive supplementary data and documentation at `http://atlas.gs.washington.edu/mouse-atac/`.

4.7 Project acknowledgments

We thank D. Prunkard and L. Gitari for exceptional assistance with flow sorting; the Neale lab for access to UK Biobank GWAS; R. Walters for advice on partitioned LDSC; W. Noble for GPU cluster access; and W. Noble, A. Adey, G. Findlay, M. Gasperini, M. Spielmann, V. Ramani, R. Chawla, and X. Qiu for valuable discussions and feedback. Funding is from the Paul G. Allen Frontiers Group (J.S. and C.T.), the W.M. Keck Foundation (C.T. and J.S.), an Alfred P. Sloan Foundation Research Fellowship (C.T.), the NIH (DP1HG007811 and R01HG006283 to J.S., DP2HD088158 to C.T., and R01GM046883 to C.M.D.). D.A.C. was supported in part by the NHLBI (T32HL007828), A.J.H. and H.A.P. were supported by NSF Graduate Research Fellowships, and W.S.D. was supported in part by the NHGRI (5T32HG000035-23). J.S. is a Howard Hughes Medical Institute Investigator.

Chapter 5: CLOSING REMARKS

While we have made enormous strides in technology development within the single-cell and genetic screening fields over the past decade, much remains to be gained both from the application of technologies in hand and future technology development. In addition to the heavily research-focused projects that I presented in Chapters 2-4, there are many other aspects of the work that would not have been obvious in the papers themselves, but nonetheless has been very important to the adoption of methods and utilization of data resources that I have worked on. Here I reflect on some of these supplementary efforts and also lay out future directions related to the work in this dissertation.

5.1 COMMODITIZATION OF SINGLE-CELL ASSAYS

Academic labs, including ours, often publish prototype methods that are not optimized with respect to performance, robustness, complexity, or labor. Labs that develop the techniques themselves are also rarely in the best position to provide detailed documentation or computational support to go along with their methods due to a lack of direct incentives. The net result is that methods must gain enough interest within the field as a whole to attract the attention of computational method/tool developers. This can often result in a ”chicken and the egg” scenario where a method is not well-supported enough to gain sufficient adoption that would otherwise lead to interest from tool developers.

As part of my graduate school training I organized an internship with 10X genomics, the first company to provide a commercial solution for droplet-based scRNA-seq. Perhaps one of the strongest qualities of 10X is their dedication to building simple, robust, performant products that are always released alongside end-to-end pipelines for processing, analyzing, and visualizing the data generated from experiments. This commitment to simplicity and support have resulted in rapid and widespread adoption of their products within both academia and industry. As an in-

tern, I was able to contribute to late stages of product development for the second version of their 3' single-cell RNA-seq assay as well as very early stages for the first version of their paired T and B-cell receptor sequencing product, giving me exposure to a wide-range of the product development cycle. I was able to contribute to assay development, software development, and novel interactive visualization tools for showing assembled T and B-cell receptor sequences along with clonotypes abundances (this was used internally throughout research and development before inspiring a new customer-facing dedicated visualization tool). I was also able to directly influence specifications and design choices for the company's first scRNA-seq visualization tool, which has since been upgraded to support several different diverse technologies. Overall, this was a tremendous opportunity to gain exposure to an early-stage company that has since played a critical role in development and adoption of single-cell methods over the past few years. 10X has gone on to release all products I contributed to along with single-cell methods for CRISPR-Cas9 guide readout, oligo-tagged antibody readouts, DNA-seq, ATAC-seq, and Spatial Transcriptomics. They continue to take on ambitious new efforts to further the state of the art and having been able to be a part of 10X, even just for a short time, had a marked impact on my graduate training.

Overall, this has shown me how influential various types of support and optimization surrounding a given method can be in determining its adoption. It has also given me a rigorous framework for thinking about technology development that has aided tremendously in thinking about my own projects in graduate school.

5.2 Data availability

While all of the studies I have been a part of during graduate school have ultimately required deposition to common repositories like GEO, this is the minimum that one could possibly do in providing primary data, processed data, and important results/methods/metadata to the community. For each of the chapters of this dissertation I have made a substantial effort to make documentation and broad data release a priority.

To promote adoption of some of the methods we have used for reading out CRISPR-based

genetic screens with single-cell RNA-seq, I set up a website providing code, data, and protocol access for the community: https://github.com/shendurelab/single-cell-ko-screens.

For our large multiplexed screen of enhancer-gene pairs in the genome we have made an effort to document important data files hosted on GEO and provide some of the key scripts for hit-calling: https://github.com/shendurelab/tafka-crisprQTL.

I have also made a website for our mouse sci-ATAC-seq atlas here: http://atlas.gs.washington.edu/mouse-atac/, and helped to set up a common data portal for a whole set of similar projects within the lab http://atlas.gs.washington.edu/hub/.

As we continue to introduce many new potential tools to the biology community, it is increasingly important to enable those without substantial domain expertise in genomics technology development to be able to use protocols, process their data, and interpret existing datasets that others have published.

5.3 VISUALIZATION TOOLS FOR SINGLE-CELL MEASUREMENTS

Throughout graduate school and my time at 10X genomics, I have developed a few interactive visualization tools for analyzing complex datasets. This is a skillset that I did not have prior to graduate school and have developed largely through teaching myself.

One such tool was developed in collaboration with Timothy Durham from Bob Waterston's lab in UW GS to simultaneously visualize data from the EPIC resource http://epic.gs.washington.edu/. This resource simultaneously draws on many different datatypes as measured throughout early *C. Elegans* embryonic development. For example, this resource contains information about cell lineage, 3-D position of cells, and protein/RNA expression data from a few different types of reporters/assays. Previously the primary way of exploring the EPIC dataset was via static images. Our tool provides a more dynamic and flexible way to monitor expression patterns over time as the embryo develops: http://epic.gs.washington.edu/epicviz/EPICViz/. This work is not yet published, although the tool is publicly available and Tim and I soon hope to prepare a short application note describing it.

As mentioned in the previous section, during my time at 10X I worked on an internal visualization tool that displayed the abundance of clonotypes along with gene and sequence difference annotations along all receptor sequences detected for each cell. This tool later served as inspiration for what has now been packaged up into a more comprehensive customer-facing tool by the 10X software group: `https://support.10xgenomics.com/single-cell-vdj/software/visualization/latest/what-is-loupe-vdj-browser`

5.4 Pipelines and tools

In addition to contributing to cellranger, 10X genomics' pipeline for processing and analyzing scRNA-seq and data from paired T and B cell receptor sequencing, I have developed many of my own automated, distributed computing pipelines for data processing. Specifically, I have developed several pipelines that are in widespread use within our lab including pipelines for single-cell RNA-seq and single-cell ATAC-seq. Each of these pipelines is designed to handle 10's of billions of reads as input across over one million cells and I have made an effort to implement very simple user interfaces for each. The scRNA-seq pipeline has been adopted as a preliminary production pipeline for the Brotman Baty Institute Advanced Technology Lab.

To facilitate development of these pipelines I have developed a general barcode correction tool, barcodeutils (`https://github.com/andrewhill157/barcodeutils`), and a workflow management tool for building distributed pipelines (easygrid, `https://github.com/andrewhill157/easygrid`).

5.5 Other publications

Beyond the work described in the dissertation (Hill et al. (2018), Gasperini et al. (2019), and Cusanovich et al. (2018)) and the work that is described in the Future directions section, I have contributed to many other publications both in and outside Genome Sciences during graduate school. Below is a list of other papers that were published either during my time in the department or that are currently under review. While these do not fit within the main arc of this dissertation, they

collectively encompass a substantial amount of time and effort on my part over the past several years.

* indicates co-first authorship

1. José L. McFaline-Figueroa, **Andrew J. Hill**, Xiaojie Qiu, Dana Jackson, Jay Shendure, Cole Trapnell. A multiplex single-cell genetic screen identifies regulatory barriers in the continuum of the epithelial-to-mesenchymal transition. **Under review**.

2. Qingbo Wang, Emma Pierce-Hoffman, Beryl B. Cummings, Konrad J. Karczewski, Jessica Alföldi, Laurent C. Francioli, Laura D. Gauthier, **Andrew J. Hill**, Anne H. O'Donnell-Luria, Genome Aggregation Database (gnomAD) Production Team, Genome Aggregation Database (gnomAD) Consortium, Daniel G. MacArthur. Landscape of multi-nucleotide variants in 125,748 human exomes and 15,708 genomes. **March 2019. bioRxiv.**

3. Junyue Cao*, Malte Spielmann*, Xiaojie Qiu, Daniel M. Ibrahim, Xingfan Huang, **Andrew J. Hill**, Fan Zhang, Stefan Mundlos, Lena Christiansen, Frank J. Steemers, Cole Trapnell, Jay Shendure. The dynamic transcriptional landscape of mammalian organogenesis at single cell resolution. **February 2019. Nature** 566,496–502.

4. Junyue Cao, Darren A. Cusanovich, Vijay Ramani, Hannah Pliner, **Andrew Hill**, Delasa Aghamirzaie, Riza Daza, Jose McFaline, Jonathan S. Packer, Lena Christiansen, Frank J. Steemers, Cole Trapnell, Jay Shendure. Joint profiling of chromatin accessibility and transcription in 15,000 single cells by combinatorial indexing. **August 2018. Science** 10.1126/science.aau0730.

5. Xiaojie Qiu, **Andrew Hill**, Jonathan Packer, Dejun Lin, Yian Ma, Cole Trapnell. Single-cell mRNA quantification and differential analysis with Census. **January 2017. Nature Methods** 14, 309–315.

6. Exome Aggregation Consortium, Monkol Lek, Konrad J Karczewski, Eric V Minikel, Kaitlin E Samocha, Eric Banks, Timothy Fennell, Anne H O'Donnell Luria, James S Ware, **Andrew J Hill**, Beryl B Cummings, Taru Tukiainen, Daniel P Birnbaum, Jack A Kosmicki, Laramie Duncan, Karol Estrada, Fengmei Zhao, James Zou, **[54 additional authors]**, Mark J Daly, Daniel G MacArthur. Combined analysis of protein-coding genetic variation in 60,706 humans. **August 2016. Nature** 536,285–291.

7. Matthew W Snyder*, Martin Kircher*, **Andrew J Hill**, Riza Daza, and Jay Shendure. Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. **January 2016. Cell**, 164(1-2), 57–68.

8. Xinxian Deng*, Wenxiu Ma*, Vijay Ramani, Andrew Hill, Fan Yang, Ferhat Ay, Joel B. Berletch, Carl Anthony Blau, Jay Shendure, Zhijun Duan, William S. Noble, and Christine M. Disteche. Bipartite structure of the inactive mouse X chromosome. **August 2015. Genome Biology**, 16:152.

5.6 FUTURE DIRECTIONS

5.6.1 Molecular readouts for genetic screens

To follow up on our work to date in using single-cell RNA-seq as a readout for CRISPR-based genetic screens, we have begun work in a number of new directions. First, we have conducted a screen knocking out putative regulators of the epithelial to mesenchymal transition (EMT) within an in-vitro model system for EMT to supplement other experimental work done in this area (McFaline-Figueroa, et al. as mentioned in the Other publications section). Second, we have begun a large-scale screen where we knock out or knock down a set of genes including all non-redundant transcription factors in mouse embryonic stem cells before forming embryoid bodies. In this case our phenotype is simply association between genotypes and bias in representation within downstream cell populations, which requires very little complexity and sequencing depth to measure (greatly reduces costs). Given this relaxed dependence on overall complexity, we are exploring the use of a combinatorial indexing-based scRNA-seq protocol that has very high throughput at the expense of reduced data quality (Cao et al., 2019). Third, I have been peripherally onvolved in a project that aims to perform targeted scRNA-seq of a small number of transcripts to enable more sensitive high-throughput screening of how individual or combinatorial perturbation of enhancers within a given TAD impact local gene expression. This will hopefully both 1) address enhancer redundancy as a major potential source of false negatives in most prior work in this field and 2) increase sensitivity by allowing targeted RT at several points along each transcript. Fourth, we are pursuing other

uses of multiplexed perturbations such as assaying synthetic lethal interactions at a genome-wide scale.

Importantly, one or more of these projects rely on combinatorial CRISPR peturbations. It is worth noting that embedding dual guide constructs within the LTR (as in CROP-seq design) would likely not result in a viable virus due to disruption of the LTR (Datlinger et al., 2017). Therefore, we will need to explore alternative approaches such as direct guide capture, which has recently been demonstrated (Replogle et al., 2018).

5.6.2 Computational prediction of enhancer-gene pairs

We hope that our efforts in Gasperini et al. (2019) will, along with several other datasets, serve as a stronger basis for predicting enhancer-gene pairs computationally from existing genomic datasets. Most recently (Fulco et al., 2019) have proposed a simple model based on H3K27ac and proximity as measured by Hi-C, although many other approaches exist and could be feasibly developed on data from high-throughput screens. We envision our approach as being complementary to approaches that use FlowFish (Fulco et al., 2019; Gasperini et al., 2019) to examine one gene at a time in that these approaches provide dense coverage of a given locus but ours provides sparse coverage of a much larger and more diverse set of genomic loci. Both types of datasets will serve as an important component of ongoing efforts such as Fulco et al. (2019) and the ENCODE consortium.

5.6.3 Single-cell ATAC-seq

The single-cell ATAC-seq protocol we used in Cusanovich et al. (2018) has several limitations that impede its widespread adoption. First, due to several technical limitations, the throughput of sci-ATAC-seq is lower than that of the equivalent sci-RNA-seq assay, making the collection of large datasets (such as those in Cusanovich et al. (2018)) very laborious. Furthermore, the two barcoding steps in this reaction are carried out via indexed transposition and indexed PCR, the first of which requires custom-loaded Tn5 enzyme that is tricky to make and maintain stable stocks of and is not a commercially available reagent (we obtain our own stocks from Illumina via a collaboration). A

diagram of this protocol is included as FIgure 1.6.

As part of our future work in this area, we have worked to develop a revised version of this assay. Briefly, rather than using custom indexed-transposons, we use commercially available Tn5 followed by phosphorylation of the loaded oligos *in-situ*. We then ligate on a barcoded oligo using a splint ligation, pool the cells, redistribute into wells, and then ligate an oligo on the other end via another splint ligation (Tn5 is loaded with two different sequences). We then pool cells again and redistribute them into a final set of wells for indexed PCR, providing three total rounds of indexing (see Figure 5.1 for a diagram of this new protocol).

We have optimized this method to achieve reduced but still reasonable performance with respect to the previous assay, but with dramatically higher throughput. For example, we have collected data from over 1 million human fetal cells across many tissues provided by the Birth Defects Research Laboratory. In addition, we have collected matched scRNA-seq data and are tackling the challenge of integration across different datatypes at scale. This project has also necessitated substantial reworking of the computational pipeline that we use to process the data given the very large number of cells and approximately 50 billions sequencing reads required thus far.

One example, of several, is that the sparsity and lower signal to noise in the assay requires revised methods for dimensionality reduction, several of which are described in an extensive blog post that I have put together: `http://andrewjohnhill.com/blog/2019/05/06/dimensionality-reductior` which I plan to prepare as a short bioRxiv post. Overall, I am very excited about what will likely be a new wave of computational methods development surrounding single-cell epigenetic assays (like scATAC-seq) as the experimental methods become more accessible.

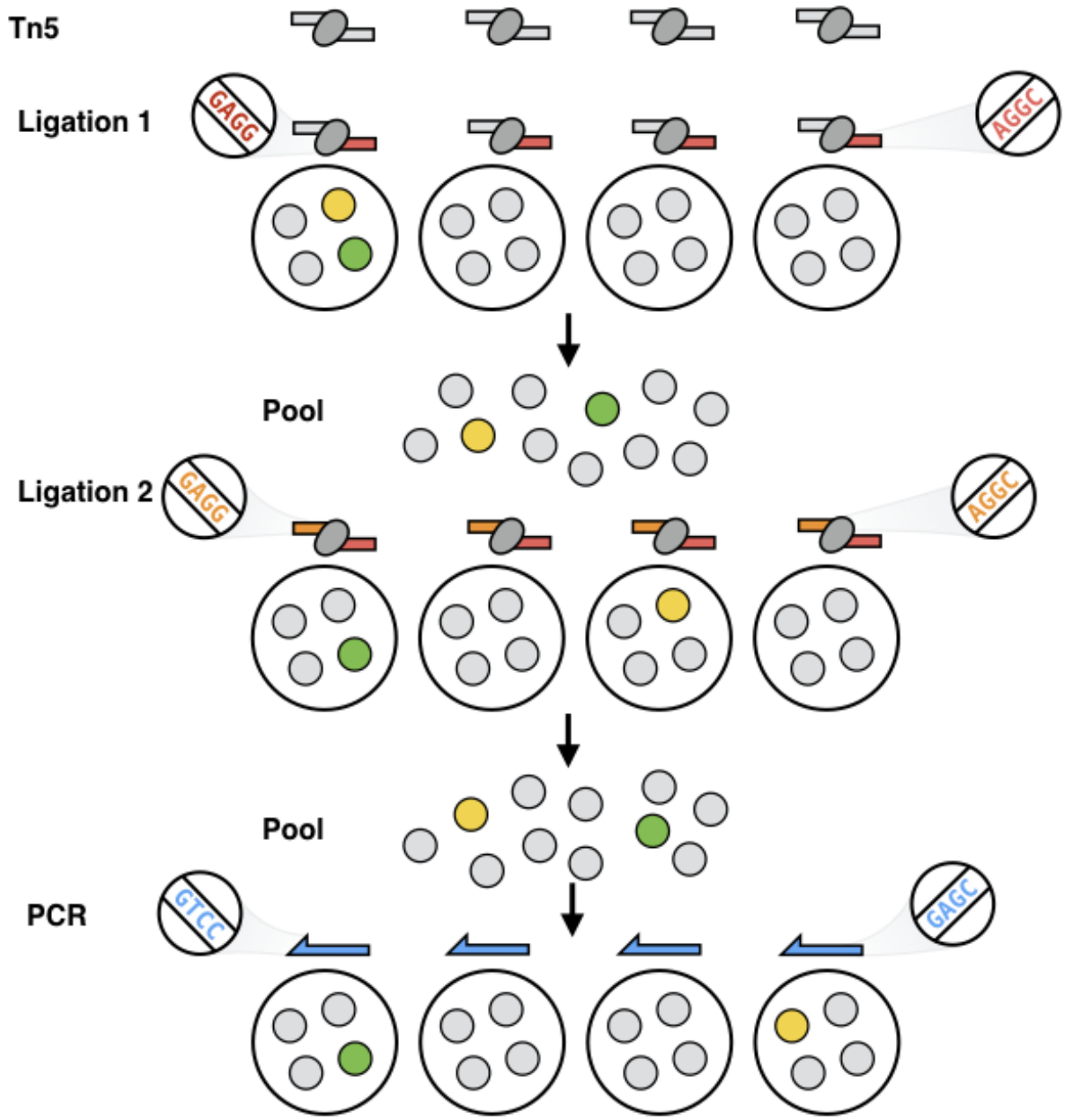


Figure 5.1: Diagram of new three-level sci-ATAC-seq protocol. Three rounds of barcoding are conducted using two rounds of splint ligation onto DNA sequences loaded on commercial Tn5 followed by an indexed PCR. In contrast, our old protocol is presented in Figure 1.6.

5.7 TOWARDS MULTI-MODAL SINGLE-CELL DATASETS

While single-cell RNA-seq or modifications of it have been the most widely adopted of single-cell genomic methods by far, there are also many other types of molecular measurements via sequencing that could in theory be ported to single-cell formats (such as ATAC-seq as we have done). These methods fall into two main categories: 1) assays that measure a small number of additional tags or specific quantities in addition to the full molecular profile and 2) assays that measure entirely different molecular profiles jointly in the same cells. As examples of the former, protocols already exist for pairing scRNA-seq with B and T-cell receptor sequences (now a 10X genomics product), pairing, scRNA-seq along with other endogenous sequences like the influenza genome in flu infected cells (Russell et al., 2018), scRNA-seq or scATAC-seq with lineage information (Raj et al., 2018; Ludwig et al., 2019), scRNA-seq with protein measurements via oligo-tagged antibodies (Peterson et al., 2017; Stoeckius et al., 2017), various sample multiplexing efforts for scRNA-seq using additional sample tags read out via scRNA-seq (Stoeckius et al., 2018; McGinnis et al., 2018; Gehring et al., 2018), and obviously work from ourselves and others surrounding the readout of genetic screens using scRNA-seq as described in Chapters 2 and 3 and recent work combining genotype readouts with scATAC-seq (Rubin et al., 2019) and microscopy (Feldman et al., 2018). Due to the technically challenging nature of genuine coassays of entirely different molecular profiles, relatively few exist in the literature, with one example being Cao et al. (2018), which develops an assay for the joint measurement of scRNA-seq and scATAC-seq. Work in this space is an ongoing area of research and these methods will require further optimization to enable their widespread adoption. Co-assays will likely be most useful/effective when one of the measured components is scRNA-seq (or something strongly correlated with gene expression) given that it is a very useful datatype for cell type identification and would facilitate easy integration with existing datasets or other coassays.

A reasonable approach in the absence of genuine coassays are approaches that aim to integrate datasets collected in parallel by converting assays like scATAC-seq to approximations of expression matrices. While the extent to which this approach will work will likely be variable across datasets

and assays, there have now been several successful attempts (including our own) at integrating scRNA-seq and scATAC-seq data (Cusanovich et al., 2018; Stuart et al., 2018; Graybuck et al., 2019; Lake et al., 2017).

It is worth noting that we certainly do not need to measure every possible quantity in every experimental context. First, each application will have its own set of measurements that are useful in addressing underlying questions. Second, it has been shown that many marks can readily be imputed given incomplete measurement of all assays across large sets of samples (Ernst and Kellis, 2015; Durham et al., 2018; Schreiber et al., 2018). This prediction task effectively depends on correlation and thus some degree of redundancy between assays. Nonetheless it will be an interesting future direction to enable several generic types of assays and learn how to prioritize given the likely upper bound on the number of molecular profiles that can be simultaneously measured. Despite these challenges, the adoption of modular and highly multi-modal experimental designs remains a very exciting future direction for the field.

5.7.1 Spatial technologies

There is broad consensus in the field that the absence of spatial context represents a substantial limitation in applications like the profiling of cell types in whole tissues. Efforts such as those in Stuart et al. (2018) and many other studies have shown that single-cell RNA-seq data can be effectively integrated with data from some spatial technologies, thereby predicting the spatial coordinates of cells based on their transcriptional similarity over genes measured by both technologies. However, there exist many potential advantages to technologies that can directly measure transcriptomes (or subsets thereof) *in-situ*. While these technologies are fairly limited in the number of labs that have been able to effectively use them and the technologies are rapidly evolving (including several commercial efforts by 10X genomics, Nanostring, and ReadCoor), it seems clear that these will play an important role in future studies where spatial information is important.

It is worth noting that there are many applications that would not necessarily benefit from spatial information, such as most applications using in-vitro systems. To my knowledge, current

efforts have been limited to measuring RNA/DNA and further technology development would be needed to provide in-situ measurements of other quantities like chromatin accessibility.

5.8 Conclusion

Much has been done in the past 5-10 years to increase the throughput and performance of single-cell technologies. While the core technology itself will continue to be refined and optimized and additional paradigms will continue to emerge, we have largely achieved a level of throughput sufficient for the vast majority of applications. In fact, many non-targeted applications are already limited equally by cell capture and sequencing costs, limiting returns on the reduction of capture costs.

Perhaps one of the most enticing potential outcomes of the ongoing wave of technology development in single-cell measurements will be truly flexible multi-modal measurements of arbitrary cell populations. Ideally, virtually any additional piece of information including entire modalities like chromatin accessibility, methylation, histone marks, transcription factor binding, spatial information, DNA sequencing, and/or targeted assays, could be combined at will to match the needs of the experiment, but there remain many technical challenges for the field to solve.

Alongside emerging technologies, computational methods for analyzing each of these modalities are still an extremely active area of research. While much work has been put towards the robust and flexible analysis of single-cell RNA-seq data, datasets approaching and exceeding one million cells still pose substantial computational challenges. As other modalities and particular applications grow in their adoption, it will be important to address the unique challenges of analyzing individual data types and data integration at scale. Co-measurement of many quantities across individual cells will also likely open up an entirely new set of challenges examining how to best make use of many (often correlated) layers of information.

Given the eventual set of potential tools at our disposal, perhaps one of our most important goals as a community should be a transition towards placing a strong focus on intelligent, rigorous, and creative experimental design. A potential pitfall of having access to so many possible mea-

surements is reliance on simply having made many measurements to establish novelty rather than showing their genuine value. If utilized to their utmost potential, single-cell technologies have an opportunity to transform our understanding and experimental capacity across numerous fields of biology.

BIBLIOGRAPHY

Adamson, B., Norman, T. M., Jost, M., Cho, M. Y., Nuñez, J. K., Chen, Y., Villalta, J. E., Gilbert, L. A., Horlbeck, M. A., Hein, M. Y., Pak, R. A., Gray, A. N., Gross, C. A., Dixit, A., Parnas, O., Regev, A. and Weissman, J. S. (2016). A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. Cell *167*, 1867–1882.e21.

Adey, A., Morrison, H. G., Asan, Xun, X., Kitzman, J. O., Turner, E. H., Stackhouse, B., MacKenzie, A. P., Caruccio, N. C., Zhang, X. and Shendure, J. (2010). Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. Genome Biol. *11*, R119.

Albert, I., Mavrich, T. N., Tomsho, L. P., Qi, J., Zanton, S. J., Schuster, S. C. and Pugh, B. F. (2007). Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome. Nature *446*, 572–576.

Amini, S., Pushkarev, D., Christiansen, L., Kostem, E., Royce, T., Turk, C., Pignatelli, N., Adey, A., Kitzman, J. O., Vijayan, K., Ronaghi, M., Shendure, J., Gunderson, K. L. and Steemers, F. J. (2014). Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. Nat. Genet. *46*, 1343–1349.

Arendt, D., Musser, J. M., Baker, C. V. H., Bergman, A., Cepko, C., Erwin, D. H., Pavlicev, M., Schlosser, G., Widder, S., Laubichler, M. D. and Wagner, G. P. (2016). The origin and evolution of cell types. Nat. Rev. Genet. *17*, 744–757.

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W. and Noble, W. S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. *37*, W202–8.

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I. and Zhao, K. (2007). High-Resolution Profiling of Histone Methylations in the Human Genome. Cell *129*, 823–837.

Bianconi, E., Piovesan, A., Facchin, F., Beraudi, A., Casadei, R., Frabetti, F., Vitale, L., Pelleri, M. C., Tassani, S., Piva, F., Perez-Amodio, S., Strippoli, P. and Canaider, S. (2013). An estimation of the number of cells in the human body. Ann. Hum. Biol. *40*, 463–471.

Bolger, A. M., Lohse, M. and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

Bray, N. L., Pimentel, H., Melsted, P. and Pachter, L. (2016). Erratum: Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. *34*, 888.

Buck, M. J. and Lieb, J. D. (2004). ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. Genomics *83*, 349–360.

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. and Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods *10*, 1213–1218.

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

Butler, A., Hoffman, P., Smibert, P., Papalexi, E. and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol. *36*, 411–420.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., McVean, G., Leslie, S., Donnelly, P. and Marchini, J. (2017). Genome-wide genetic data on 500,000 UK biobank participants.

Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A. and Rinn, J. L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. *25*, 1915–1927.

Canté-Barrett, K., Pieters, R. and Meijerink, J. P. P. (2014). Myocyte enhancer factor 2C in hematopoiesis and leukemia. Oncogene *33*, 403–410.

Canver, M. C., Smith, E. C., Sher, F., Pinello, L., Sanjana, N. E., Shalem, O., Chen, D. D., Schupp, P. G., Vinjamur, D. S., Garcia, S. P., Luc, S., Kurita, R., Nakamura, Y., Fujiwara, Y., Maeda, T., Yuan, G.-C., Zhang, F., Orkin, S. H. and Bauer, D. E. (2015). BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. Nature *527*, 192–197.

Cao, F., Fang, Y., Tan, H. K., Goh, Y., Choy, J. Y. H., Koh, B. T. H., Hao Tan, J., Bertin, N., Ramadass, A., Hunter, E., Green, J., Salter, M., Akoulitchev, A., Wang, W., Chng, W. J., Tenen, D. G. and Fullwood, M. J. (2017). Super-Enhancers and Broad H3K4me3 Domains Form Complex Gene Regulatory Circuits Involving Chromatin Interactions. Sci. Rep. *7*, 2186.

Cao, J., Cusanovich, D. A., Ramani, V., Aghamirzaie, D., Pliner, H. A., Hill, A. J., Daza, R. M., McFaline-Figueroa, J. L., Packer, J. S., Christiansen, L., Steemers, F. J., Adey, A. C., Trapnell, C. and Shendure, J. (2018). Joint profiling of chromatin accessibility and gene expression in thousands of single cells. Science *361*, 1380–1385.

Cao, J., Packer, J. S., Ramani, V., Cusanovich, D. A., Huynh, C., Daza, R., Qiu, X., Lee, C., Furlan, S. N., Steemers, F. J., Adey, A., Waterston, R. H., Trapnell, C. and Shendure, J. (2017). Comprehensive single-cell transcriptional profiling of a multicellular organism. Science *357*, 661–667.

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D. M., Hill, A. J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F. J., Trapnell, C. and Shendure, J. (2019). The single-cell transcriptional landscape of mammalian organogenesis. Nature *566*, 496–502.

Chen, B., Gilbert, L. A., Cimini, B. A., Schnitzbauer, J., Zhang, W., Li, G.-W., Park, J., Blackburn, E. H., Weissman, J. S., Qi, L. S. and Huang, B. (2013). Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. Cell *155*, 1479–1491.

Choi, H. M. T., Schwarzkopf, M., Fornace, M. E., Acharya, A., Artavanis, G., Stegmaier, J., Cunha, A. and Pierce, N. A. (2018). Third-generation in situ hybridization chain reaction: multiplexed, quantitative, sensitive, versatile, robust. Development *145*.

Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P. D., Wu, X., Jiang, W., Marraffini, L. A. and Zhang, F. (2013). Multiplex genome engineering using CRISPR/Cas systems. Science *339*, 819–823.

Contente, A., Dittmer, A., Koch, M. C., Roth, J. and Dobbelstein, M. (2002). A polymorphic microsatellite that mediates induction of PIG3 by p53. Nat. Genet. *30*, 315–320.

Cusanovich, D. A., Daza, R., Adey, A., Pliner, H. A., Christiansen, L., Gunderson, K. L., Steemers, F. J., Trapnell, C. and Shendure, J. (2015). Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. Science *348*, 910–914.

Cusanovich, D. A., Hill, A. J., Aghamirzaie, D., Daza, R. M., Pliner, H. A., Berletch, J. B., Filippova, G. N., Huang, X., Christiansen, L., DeWitt, W. S., Lee, C., Regalado, S. G., Read, D. F., Steemers, F. J., Disteche, C. M., Trapnell, C. and Shendure, J. (2018). A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. Cell *174*, 1309–1324.e18.

Cusanovich, D. A., Reddington, J. P., Garfield, D. A., Daza, R., Marco-Ferreres, R., Christiansen, L., Qiu, X., Steemers, F., Trapnell, C., Shendure, J. and Furlong, E. E. M. (2017). The cis-regulatory dynamics of embryonic development at single cell resolution.

Datlinger, P., Rendeiro, A. F., Schmidl, C., Krausgruber, T., Traxler, P., Klughammer, J., Schuster, L. C., Kuchler, A., Alpar, D. and Bock, C. (2017). Pooled CRISPR screening with single-cell transcriptome readout. Nat. Methods *14*, 297–301.

Debnath, J., Muthuswamy, S. K. and Brugge, J. S. (2003). Morphogenesis and oncogenesis of MCF-10A mammary epithelial acini grown in three-dimensional basement membrane cultures. Methods *30*, 256–268.

Der, E., Ranabothu, S., Suryawanshi, H., Akat, K. M., Clancy, R., Morozov, P., Kustagi, M., Czuppa, M., Izmirly, P., Belmont, H. M., Wang, T., Jordan, N., Bornkamp, N., Nwaukoni, J., Martinez, J., Goilav, B., Buyon, J. P., Tuschl, T. and Putterman, C. (2017). Single cell RNA sequencing to dissect the molecular heterogeneity in lupus nephritis. JCI Insight *2*.

Diao, Y., Fang, R., Li, B., Meng, Z., Yu, J., Qiu, Y., Lin, K. C., Huang, H., Liu, T., Marina, R. J., Jung, I., Shen, Y., Guan, K.-L. and Ren, B. (2017). A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. Nat. Methods *14*, 629–635.

Diao, Y., Li, B., Meng, Z., Jung, I., Lee, A. Y., Dixon, J., Maliskova, L., Guan, K.-L., Shen, Y. and Ren, B. (2016). A new class of temporarily phenotypic enhancers identified by CRISPR/Cas9-mediated genetic screening. Genome Res. *26*, 397–405.

Dixit, A. (2016). Correcting Chimeric Crosstalk in Single Cell RNA-seq Experiments.

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-Arnon, L., Marjanovic, N. D., Dionne, D., Burks, T., Raychowdhury, R., Adamson, B., Norman, T. M., Lander, E. S., Weissman, J. S., Friedman, N. and Regev, A. (2016). Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. Cell *167*, 1853–1866.e17.

Dostie, J., Richmond, T. A., Arnaout, R. A., Selzer, R. R., Lee, W. L., Honan, T. A., Rubio, E. D., Krumm, A., Lamb, J., Nusbaum, C., Green, R. D. and Dekker, J. (2006). Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res. *16*, 1299–1309.

Durham, T. J., Libbrecht, M. W., Howbert, J. J., Bilmes, J. and Noble, W. S. (2018). PREDICTD PaRallel Epigenomics Data Imputation with Cloud-based Tensor Decomposition. Nature Communications *9*.

Eisenberg, E. and Levanon, E. Y. (2013). Human housekeeping genes, revisited. Trends Genet. *29*, 569–574.

el Deiry, W. S., Tokino, T., Velculescu, V. E., Levy, D. B., Parsons, R., Trent, J. M., Lin, D., Mercer, W. E., Kinzler, K. W. and Vogelstein, B. (1993). WAF1, a potential mediator of p53 tumor suppression. Cell *75*, 817–825.

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

Ernst, J. and Kellis, M. (2015). Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. Nature Biotechnology *33*, 364–376.

Feldman, D., Singh, A., Schmid-Burgk, J. L., Mezger, A., Garrity, A. J., Carlson, R. J., Zhang, F. and Blainey, P. (2018). Pooled optical screens in human cells. bioRxiv .

Fincher, C. T., Wurtzel, O., de Hoog, T., Kravarik, K. M. and Reddien, P. W. (2018). Cell type transcriptome atlas for the planarian Schmidtea mediterranea. Science .

Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K., Ripke, S., Day, F. R., ReproGen Consortium, Schizophrenia Working

Group of the Psychiatric Genomics Consortium, RACI Consortium, Purcell, S., Stahl, E., Lindstrom, S., Perry, J. R. B., Okada, Y., Raychaudhuri, S., Daly, M. J., Patterson, N., Neale, B. M. and Price, A. L. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. *47*, 1228–1235.

Frieda, K. L., Linton, J. M., Hormoz, S., Choi, J., Chow, K.-H. K., Singer, Z. S., Budde, M. W., Elowitz, M. B. and Cai, L. (2017). Synthetic recording and in situ readout of lineage information in single cells. Nature *541*, 107–111.

Fulco, C. P., Munschauer, M., Anyoha, R., Munson, G., Grossman, S. R., Perez, E. M., Kane, M., Cleary, B., Lander, E. S. and Engreitz, J. M. (2016). Systematic mapping of functional enhancer-promoter connections with CRISPR interference. Science *354*, 769–773.

Fulco, C. P., Nasser, J., Jones, T. R., Munson, G., Bergman, D. T., Subramanian, V., Grossman, S. R., Anyoha, R., Patwardhan, T. A., Nguyen, T. H., Kane, M., Doughty, B., Perez, E. M., Durand, N. C., Stamenova, E. K., Aiden, E. L., Lander, E. S. and Engreitz, J. M. (2019). Activity-by-contact model of enhancer specificity from thousands of CRISPR perturbations.

Gambone, J. E., Dusaban, S. S., Loperena, R., Nakata, Y. and Shetzline, S. E. (2011). The c-Myb target gene neuromedin U functions as a novel cofactor during the early stages of erythropoiesis. Blood *117*, 5733–5743.

Ganapathi, M., Srivastava, P., Das Sutar, S. K., Kumar, K., Dasgupta, D., Pal Singh, G., Brahmachari, V. and Brahmachari, S. K. (2005). Comparative analysis of chromatin landscape in regulatory regions of human housekeeping and tissue specific genes. BMC Bioinformatics *6*, 126.

Gasperini, M., Findlay, G. M., McKenna, A., Milbank, J. H., Lee, C., Zhang, M. D., Cusanovich, D. A. and Shendure, J. (2017). CRISPR/Cas9-Mediated Scanning for Regulatory Elements Required for HPRT1 Expression via Thousands of Large, Programmed Genomic Deletions. Am. J. Hum. Genet. *101*, 192–205.

Gasperini, M., Hill, A. J., McFaline-Figueroa, J. L., Martin, B., Kim, S., Zhang, M. D., Jackson, D., Leith, A., Schreiber, J., Noble, W. S., Trapnell, C., Ahituv, N. and Shendure, J. (2019). A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. Cell *176*, 377–390.e19.

Gasperini, M., Starita, L. and Shendure, J. (2016). The power of multiplexed functional analysis of genetic variants. Nature Protocols *11*, 1782–1787.

Gehring, J., Park, J. H., Chen, S., Thomson, M. and Pachter, L. (2018). Highly Multiplexed Single-Cell RNA-seq for Defining Cell Population and Transcriptional Spaces. bioRxiv .

Gierahn, T. M., II, M. H. W., Hughes, T. K., Bryson, B. D., Butler, A., Satija, R., Fortune, S., Love, J. C. and Shalek, A. (2017). Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. Protocol Exchange .

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

Graybuck, L. T., Sedeño-Cortés, A., Nguyen, T. N., Walker, M., Szelenyi, E., Lenz, G., Sieverts, L., Kim, T. K., Garren, E., Kalmbach, B., Yao, S., Mortrud, M., Mich, J., Goldy, J., Smith, K. A., Dee, N., Yao, Z., Cetin, A., Levi, B. P., Lein, E., Ting, J., Zeng, H., Daigle, T. and Tasic, B. (2019). Prospective, brain-wide labeling of neuronal subclasses with enhancer-driven AAVs. bioRxiv .

Guo, G., Huss, M., Tong, G. Q., Wang, C., Sun, L. L., Clarke, N. D. and Robson, P. (2010). Resolution of Cell Fate Decisions Revealed by Single-Cell Gene Expression Analysis from Zygote to Blastocyst. Developmental Cell *18*, 675–685.

Haghverdi, L., Lun, A. T. L., Morgan, M. D. and Marioni, J. C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat. Biotechnol. .

Han, K., Jeng, E. E., Hess, G. T., Morgens, D. W., Li, A. and Bassik, M. C. (2017). Synergistic drug combinations for cancer identified in a CRISPR screen for pairwise genetic interactions. Nat. Biotechnol. *35*, 463–474.

Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., Saadatpour, A., Zhou, Z., Chen, H., Ye, F., Huang, D., Xu, Y., Huang, W., Jiang, M., Jiang, X., Mao, J., Chen, Y., Lu, C., Xie, J., Fang, Q., Wang, Y., Yue, R., Li, T., Huang, H., Orkin, S. H., Yuan, G.-C., Chen, M. and Guo, G. (2018). Mapping the Mouse Cell Atlas by Microwell-Seq. Cell *172*, 1091–1107.e17.

Hesselberth, J. R., Chen, X., Zhang, Z., Sabo, P. J., Sandstrom, R., Reynolds, A. P., Thurman, R. E., Neph, S., Kuehn, M. S., Noble, W. S., Fields, S. and Stamatoyannopoulos, J. A. (2009). Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. Nat. Methods *6*, 283–289.

Hill, A. J., McFaline-Figueroa, J. L., Starita, L. M., Gasperini, M. J., Matreyek, K. A., Packer, J., Jackson, D., Shendure, J. and Trapnell, C. (2018). On the design of CRISPR-based single-cell molecular screens. Nat. Methods *15*, 271–274.

Hindson, B. J., Ness, K. D., Masquelier, D. A., Belgrader, P., Heredia, N. J., Makarewicz, A. J., Bright, I. J., Lucero, M. Y., Hiddessen, A. L., Legler, T. C., Kitano, T. K., Hodel, M. R., Petersen, J. F., Wyatt, P. W., Steenblock, E. R., Shah, P. H., Bousse, L. J., Troup, C. B., Mellen, J. C., Wittmann, D. K., Erndt, N. G., Cauley, T. H., Koehler, R. T., So, A. P., Dube, S., Rose, K. A., Montesclaros, L., Wang, S., Stumbo, D. P., Hodges, S. P., Romine, S., Milanovich, F. P., White, H. E., Regan, J. F., Karlin-Neumann, G. A., Hindson, C. M., Saxonov, S. and Colston, B. W.

(2011). High-Throughput Droplet Digital PCR System for Absolute Quantitation of DNA Copy Number. Analytical Chemistry *83*, 8604–8610.

Hong, J.-W., Hendrix, D. A. and Levine, M. S. (2008). Shadow enhancers as a source of evolutionary novelty. Science *321*, 1314.

Horlbeck, M. A., Gilbert, L. A., Villalta, J. E., Adamson, B., Pak, R. A., Chen, Y., Fields, A. P., Park, C. Y., Corn, J. E., Kampmann, M. and Weissman, J. S. (2016). Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. Elife *5*.

Hosoya, T., Maillard, I. and Engel, J. D. (2010). From the cradle to the grave: activities of GATA-3 throughout T-cell development and differentiation. Immunol. Rev. *238*, 110–125.

Islam, S., Kjallquist, U., Moliner, A., Zajac, P., Fan, J.-B., Lonnerberg, P. and Linnarsson, S. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Research *21*, 1160–1167.

Jaitin, D. A., Weiner, A., Yofe, I., Lara-Astiaso, D., Keren-Shaul, H., David, E., Salame, T. M., Tanay, A., van Oudenaarden, A. and Amit, I. (2016). Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq. Cell *167*, 1883–1896.e15.

Jetzt, A. E., Yu, H., Klarmann, G. J., Ron, Y., Preston, B. D. and Dougherty, J. P. (2000). High rate of recombination throughout the human immunodeficiency virus type 1 genome. J. Virol. *74*, 1234–1240.

Johnson, D. S., Mortazavi, A., Myers, R. M. and Wold, B. (2007). Genome-Wide Mapping of in Vivo Protein-DNA Interactions. Science *316*, 1497–1502.

Karaiskos, N., Wahle, P., Alles, J., Boltengagen, A., Ayoub, S., Kipar, C., Kocks, C., Rajewsky, N. and Zinzen, R. P. (2017). The Drosophila embryo at single-cell transcriptome resolution. Science *358*, 194–199.

Kelley, D. R., Snoek, J. and Rinn, J. L. (2016). Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res. *26*, 990–999.

Klann, T. S., Black, J. B., Chellappan, M., Safi, A., Song, L., Hilton, I. B., Crawford, G. E., Reddy, T. E. and Gersbach, C. A. (2017). CRISPR-Cas9 epigenome editing enables high-throughput screening for functional regulatory elements in the human genome. Nat. Biotechnol. *35*, 561–568.

Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D. A. and Kirschner, M. W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell *161*, 1187–1201.

Korkmaz, G., Lopes, R., Ugalde, A. P., Nevedomskaya, E., Han, R., Myacheva, K., Zwart, W., Elkon, R. and Agami, R. (2016). Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. Nat. Biotechnol. *34*, 192–198.

Kuhn, R. M., Haussler, D. and Kent, W. J. (2013). The UCSC genome browser and associated tools. Brief. Bioinform. *14*, 144–161.

Kulakovskiy, I. V., Medvedeva, Y. A., Schaefer, U., Kasianov, A. S., Vorontsov, I. E., Bajic, V. B. and Makeev, V. J. (2013). HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. Nucleic Acids Res. *41*, D195–202.

Kulakovskiy, I. V., Vorontsov, I. E., Yevshin, I. S., Soboleva, A. V., Kasianov, A. S., Ashoor, H., Ba-Alawi, W., Bajic, V. B., Medvedeva, Y. A., Kolpakov, F. A. and Makeev, V. J. (2016). HOCOMOCO: expansion and enhancement of the collection of transcription factor binding sites models. Nucleic Acids Res. *44*, D116–25.

Lake, B. B., Ai, R., Kaeser, G. E., Salathia, N. S., Yung, Y. C., Liu, R., Wildberg, A., Gao, D., Fung, H.-L., Chen, S., Vijayaraghavan, R., Wong, J., Chen, A., Sheng, X., Kaper, F., Shen, R., Ronaghi, M., Fan, J.-B., Wang, W., Chun, J. and Zhang, K. (2016). Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. Science *352*, 1586–1590.

Lake, B. B., Chen, S., Sos, B. C., Fan, J., Kaeser, G. E., Yung, Y. C., Duong, T. E., Gao, D., Chun, J., Kharchenko, P. V. and Zhang, K. (2017). Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. Nature Biotechnology *36*, 70–80.

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

Lara-Astiaso, D., Weiner, A., Lorenzo-Vivas, E., Zaretsky, I., Jaitin, D. A., David, E., Keren-Shaul, H., Mildner, A., Winter, D., Jung, S., Friedman, N. and Amit, I. (2014). Immunogenetics. Chromatin state dynamics during blood formation. Science *345*, 943–949.

Lavin, Y., Winter, D., Blecher-Gonen, R., David, E., Keren-Shaul, H., Merad, M., Jung, S. and Amit, I. (2014). Tissue-resident macrophage enhancer landscapes are shaped by the local microenvironment. Cell *159*, 1312–1326.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Lin, Y., Ghazanfar, S., Strbenac, D., Wang, A., Patrick, E., Speed, T., Yang, J. and Yang, P. (2017). Housekeeping genes, revisited at the single-cell level. BioRxiv , 229815.

Ludwig, L. S., Lareau, C. A., Ulirsch, J. C., Christian, E., Muus, C., Li, L. H., Pelka, K., Ge, W., Oren, Y., Brack, A., Law, T., Rodman, C., Chen, J. H., Boland, G. M., Hacohen, N., Rozenblatt-Rosen, O., Aryee, M. J., Buenrostro, J. D., Regev, A. and Sankaran, V. G. (2019). Lineage Tracing in Humans Enabled by Mitochondrial Mutations and Single-Cell Genomics. Cell *176*, 1325–1339.e22.

Lukoszek, R., Mueller-Roeber, B. and Ignatova, Z. (2013). Interplay between polymerase II- and polymerase III-assisted expression of overlapping genes. FEBS Lett. *587*, 3692–3695.

Luo, C., Keown, C. L., Kurihara, L., Zhou, J., He, Y., Li, J., Castanon, R., Lucero, J., Nery, J. R., Sandoval, J. P., Bui, B., Sejnowski, T. J., Harkins, T. T., Mukamel, E. A., Behrens, M. M. and Ecker, J. R. (2017). Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. Science *357*, 600–604.

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., Pendlington, Z. M., Welter, D., Burdett, T., Hindorff, L., Flicek, P., Cunningham, F. and Parkinson, H. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. *45*, D896–D901.

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., Trombetta, J. J., Weitz, D. A., Sanes, J. R., Shalek, A. K., Regev, A. and McCarroll, S. A. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell *161*, 1202–1214.

Matreyek, K. A., Starita, L. M., Stephany, J. J., Martin, B., Chiasson, M. A., Gray, V. E., Kircher, M., Khechaduri, A., Dines, J. N., Hause, R. J., Bhatia, S., Evans, W. E., Relling, M. V., Yang, W., Shendure, J. and Fowler, D. M. (2018). Multiplex assessment of protein variant abundance by massively parallel sequencing. Nature Genetics *50*, 874–882.

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutyavin, T., Stehling-Sun, S., Johnson, A. K., Canfield, T. K., Giste, E., Diegel, M., Bates, D., Hansen, R. S., Neph, S., Sabo, P. J., Heimfeld, S., Raubitschek, A., Ziegler, S., Cotsapas, C., Sotoodehnia, N., Glass, I., Sunyaev, S. R., Kaul, R. and Stamatoyannopoulos, J. A. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190–1195.

McGinnis, C. S., Patterson, D. M., Winkler, J., Hein, M. Y., Srivastava, V., Conrad, D. N., Murrow, L. M., Weissman, J. S., Werb, Z., Chow, E. D. and Gartner, Z. J. (2018). MULTI-seq: Scalable sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. bioRxiv .

McKenna, A., Findlay, G. M., Gagnon, J. A., Horwitz, M. S., Schier, A. F. and Shendure, J. (2016). Whole-organism lineage tracing by combinatorial and cumulative genome editing. Science *353*, aaf7907.

McKenna, A. and Shendure, J. (2018). FlashFry: a fast and flexible tool for large-scale CRISPR target design. BMC Biol. *16*, 74.

Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.-K., Koche, R. P., Lee, W., Mendenhall, E., O'Donovan, A., Presser, A., Russ, C., Xie, X., Meissner, A., Wernig, M., Jaenisch, R., Nusbaum, C., Lander, E. S. and Bernstein, B. E. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature *448*, 553–560.

Mockler, T. C. and Ecker, J. R. (2005). Applications of DNA tiling arrays for whole-genome analysis. Genomics *85*, 1–15.

Mohr, S. E., Smith, J. A., Shamu, C. E., Neumüller, R. A. and Perrimon, N. (2014). RNAi screening comes of age: improved techniques and complementary approaches. Nat. Rev. Mol. Cell Biol. *15*, 591–600.

Morley, M., Molony, C. M., Weber, T. M., Devlin, J. L., Ewens, K. G., Spielman, R. S. and Cheung, V. G. (2004). Genetic analysis of genome-wide variation in human gene expression. Nature *430*, 743–747.

Mulqueen, R. M., Pokholok, D., Norberg, S., Fields, A. J., Sun, D., Torkenczy, K. A., Shendure, J., Trapnell, C., O'Roak, B. J., Xia, Z., Steemers, F. J. and Adey, A. C. (2017). Scalable and efficient single-cell DNA methylation sequencing by combinatorial indexing.

Nikolaitchik, O. A., Dilley, K. A., Fu, W., Gorelick, R. J., Tai, S.-H. S., Soheilian, F., Ptak, R. G., Nagashima, K., Pathak, V. K. and Hu, W.-S. (2013). Dimeric RNA recognition regulates HIV-1 genome packaging. PLoS Pathog. *9*, e1003249.

Noordermeer, D. and de Laat, W. (2008). Joining the loops: beta-globin gene regulation. IUBMB Life *60*, 824–833.

Olins, A. L., Zwerger, M., Herrmann, H., Zentgraf, H., Simon, A. J., Monestier, M. and Olins, D. E. (2008). The human granulocyte nucleus: Unusual nuclear envelope and heterochromatin composition. Eur. J. Cell Biol. *87*, 279–290.

Park, J., Shrestha, R., Qiu, C., Kondo, A., Huang, S., Werth, M., Li, M., Barasch, J. and Susztak, K. (2018). Comprehensive single cell RNAseq analysis of the kidney reveals novel cell types and unexpected cell plasticity. Science .

Partanen, T. A., Arola, J., Saaristo, A., Jussila, L., Ora, A., Miettinen, M., Stacker, S. A., Achen, M. G. and Alitalo, K. (2000). VEGF-C and VEGF-D expression in neuroendocrine cells and their receptor, VEGFR-3, in fenestrated blood vessels in human tissues. FASEB J. *14*, 2087–2096.

Peterson, V. M., Zhang, K. X., Kumar, N., Wong, J., Li, L., Wilson, D. C., Moore, R., McClanahan, T. K., Sadekova, S. and Klappenbach, J. A. (2017). Multiplexed quantification of proteins and transcripts in single cells. Nature Biotechnology *35*, 936–939.

Pickrell, J. K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. Am. J. Hum. Genet. *94*, 559–573.

Pimentel, H., Bray, N. L., Puente, S., Melsted, P. and Pachter, L. (2017). Differential analysis of RNA-seq incorporating quantification uncertainty. Nat. Methods *14*, 687–690.

Plass, M., Solana, J., Wolf, F. A., Ayoub, S., Misios, A., Glažar, P., Obermayer, B., Theis, F. J., Kocks, C. and Rajewsky, N. (2018). Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. Science .

Pliner, H. A., Packer, J. S., McFaline-Figueroa, J. L., Cusanovich, D. A., Daza, R. M., Aghamirzaie, D., Srivatsan, S., Qiu, X., Jackson, D., Minkina, A., Adey, A. C., Steemers, F. J., Shendure, J. and Trapnell, C. (2018). Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. Molecular Cell *71*, 858–871.e8.

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. *20*, 110–121.

Preissl, S., Fang, R., Huang, H., Zhao, Y., Raviram, R., Gorkin, D. U., Zhang, Y., Sos, B. C., Afzal, V., Dickel, D. E., Kuan, S., Visel, A., Pennacchio, L. A., Zhang, K. and Ren, B. (2018). Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. Nature Neuroscience *21*, 432–439.

Qiu, X., Hill, A., Packer, J., Lin, D., Ma, Y.-A. and Trapnell, C. (2017a). Single-cell mRNA quantification and differential analysis with Census. Nat. Methods *14*, 309–315.

Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H. A. and Trapnell, C. (2017b). Reversed graph embedding resolves complex single-cell trajectories. Nat. Methods *14*, 979–982.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Raj, B., Wagner, D. E., McKenna, A., Pandey, S., Klein, A. M., Shendure, J., Gagnon, J. A. and Schier, A. F. (2018). Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. Nature Biotechnology *36*, 442–450.

Rajagopal, N., Srinivasan, S., Kooshesh, K., Guo, Y., Edwards, M. D., Banerjee, B., Syed, T., Emons, B. J. M., Gifford, D. K. and Sherwood, R. I. (2016). High-throughput mapping of regulatory DNA. Nat. Biotechnol. *34*, 167–174.

Ramani, V., Deng, X., Qiu, R., Gunderson, K. L., Steemers, F. J., Disteche, C. M., Noble, W. S., Duan, Z. and Shendure, J. (2017). Massively multiplex single-cell Hi-C. Nat. Methods *14*, 263–266.

Ramírez, F., Ryan, D. P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., Heyne, S., Dündar, F. and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. *44*, W160–5.

Ran, F. A., Hsu, P. D., Wright, J., Agarwala, V., Scott, D. A. and Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. Nat. Protoc. *8*, 2281–2308.

Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S. and Aiden, E. L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665–1680.

Rashid, A. J., Cole, C. J. and Josselyn, S. A. (2014). Emerging roles for MEF2 transcription factors in memory. Genes Brain Behav. *13*, 118–125.

Replogle, J. M., Xu, A., Norman, T. M., Meer, E. J., Terry, J. M., Riordan, D., Srinivas, N., Mikkelsen, T. S., Weissman, J. S. and Adamson, B. (2018). Direct capture of CRISPR guides enables scalable, multiplexed, and multi-omic perturb-seq.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., Wu, Y.-C., Pfenning, A. R., Wang, X., Claussnitzer, M., Liu, Y., Coarfa, C., Harris, R. A., Shoresh, N., Epstein, C. B., Gjoneska, E., Leung, D., Xie, W., Hawkins, R. D., Lister, R., Hong, C., Gascard, P., Mungall, A. J., Moore, R., Chuah, E., Tam, A., Canfield, T. K., Hansen, R. S., Kaul, R., Sabo, P. J., Bansal, M. S., Carles, A., Dixon, J. R., Farh, K.-H., Feizi, S., Karlic, R., Kim, A.-R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T. R., Neph, S. J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R. C., Siebenthall, K. T., Sinnott-Armstrong, N. A., Stevens, M., Thurman, R. E., Wu, J., Zhang, B., Zhou, X., Beaudet, A. E., Boyer, L. A., De Jager, P. L., Farnham, P. J., Fisher, S. J., Haussler, D., Jones, S. J. M., Li, W., Marra, M. A., McManus, M. T., Sunyaev, S., Thomson, J. A., Tlsty, T. D., Tsai, L.-H., Wang, W., Waterland, R. A., Zhang, M. Q., Chadwick, L. H., Bernstein, B. E., Costello, J. F., Ecker, J. R., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos, J. A., Wang, T. and Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.

Ross, M. H. and Pawlina, W. (1979). Histology: A Text and Atlas, with Correlated Cell and Molecular Biology, 6th Edition. Lippincott.

Rubin, A. J., Parker, K. R., Satpathy, A. T., Qi, Y., Wu, B., Ong, A. J., Mumbach, M. R., Ji, A. L., Kim, D. S., Cho, S. W., Zarnegar, B. J., Greenleaf, W. J., Chang, H. Y. and Khavari, P. A.

(2019). Coupled Single-Cell CRISPR Screening and Epigenomic Profiling Reveals Causal Gene Regulatory Networks. Cell *176*, 361–376.e17.

Russell, A. B., Trapnell, C. and Bloom, J. D. (2018). Extreme heterogeneity of influenza virus infection in single cells. eLife *7*.

Sack, L. M., Davoli, T., Xu, Q., Li, M. Z. and Elledge, S. J. (2016). Sources of Error in Mammalian Genetic Screens. G3 *6*, 2781–2790.

Sanjana, N. E., Shalem, O. and Zhang, F. (2014). Improved vectors and genome-wide libraries for CRISPR screening. Nat. Methods *11*, 783–784.

Sanjana, N. E., Wright, J., Zheng, K., Shalem, O., Fontanillas, P., Joung, J., Cheng, C., Regev, A. and Zhang, F. (2016). High-resolution interrogation of functional elements in the noncoding genome. Science *353*, 1545–1549.

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. Nat. Biotechnol. *33*, 495–502.

Schlub, T. E., Smyth, R. P., Grimm, A. J., Mak, J. and Davenport, M. P. (2010). Accurately measuring recombination between closely related HIV-1 genomes. PLoS Comput. Biol. *6*, e1000766.

Schmitt, A. D., Hu, M., Jung, I., Xu, Z., Qiu, Y., Tan, C. L., Li, Y., Lin, S., Lin, Y., Barr, C. L. and Ren, B. (2016). A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. Cell Rep. *17*, 2042–2059.

Schreiber, J., Durham, T., Bilmes, J. and Noble, W. S. (2018). Multi-scale deep tensor factorization learns a latent representation of the human epigenome. bioRxiv .

Scrucca, L., Fop, M., Murphy, T. B. and Raftery, A. E. (2016). mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. R J. *8*, 289–317.

Shalem, O., Sanjana, N. E. and Zhang, F. (2015). High-throughput functional genomics using CRISPR–Cas9. Nat. Rev. Genet. *16*, 299–311.

Smith, C. L., Goldsmith, C.-A. W. and Eppig, J. T. (2005). The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. Genome Biol. *6*, R7.

Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P. K., Swerdlow, H., Satija, R. and Smibert, P. (2017). Simultaneous epitope and transcriptome measurement in single cells. Nature Methods *14*, 865–868.

Stoeckius, M., Zheng, S., Houck-Loomis, B., Hao, S., Yeung, B. Z., Mauck, W. M., Smibert, P. and Satija, R. (2018). Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. Genome Biology *19*.

Stranger, B. E., Montgomery, S. B., Dimas, A. S., Parts, L., Stegle, O., Ingle, C. E., Sekowska, M., Smith, G. D., Evans, D., Gutierrez-Arcelus, M., Price, A., Raj, T., Nisbett, J., Nica, A. C., Beazley, C., Durbin, R., Deloukas, P. and Dermitzakis, E. T. (2012). Patterns of cis regulatory variation in diverse human populations. PLoS Genet. *8*, e1002639.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., Stoeckius, M., Smibert, P. and Satija, R. (2018). Comprehensive integration of single cell data.

Svensson, V., Vento-Tormo, R. and Teichmann, S. A. (2018). Exponential scaling of single-cell RNA-seq in the past decade. Nature Protocols *13*, 599–604.

Tanabe, O., McPhee, D., Kobayashi, S., Shen, Y., Brandt, W., Jiang, X., Campbell, A. D., Chen, Y.-T., Chang, C. S., Yamamoto, M., Tanimoto, K. and Engel, J. D. (2007). Embryonic and fetal beta-globin gene repression by the orphan nuclear receptors, TR2 and TR4. EMBO J. *26*, 2295–2306.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., Lao, K. and Surani, M. A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. Nature Methods *6*, 377–382.

Taylor, A. C., Murfee, W. L. and Peirce, S. M. (2007). EphB4 expression along adult rat microvascular networks: EphB4 is more than a venous specific marker. Microcirculation *14*, 253–267.

Thakore, P. I., D'Ippolito, A. M., Song, L., Safi, A., Shivakumar, N. K., Kabadi, A. M., Reddy, T. E., Crawford, G. E. and Gersbach, C. A. (2015). Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. Nat. Methods *12*, 1143–1149.

The Tabula Muris Consortium, Quake, S. R., Wyss-Coray, T. and Darmanis, S. (2017). Transcriptomic characterization of 20 organs and tissues from mouse at single cell resolution creates a tabula muris.

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T. K., Diegel, M., Dunn, D., Ebersol, A. K., Frum, T., Giste, E., Johnson, A. K., Johnson, E. M., Kutyavin, T., Lajoie, B., Lee, B.-K., Lee, K., London, D., Lotakis, D., Neph, S., Neri, F., Nguyen, E. D., Qu, H., Reynolds, A. P., Roach, V., Safi, A., Sanchez, M. E., Sanyal, A., Shafer, A., Simon, J. M., Song, L., Vong, S., Weaver, M., Yan, Y., Zhang, Z., Zhang, Z., Lenhard, B., Tewari, M., Dorschner, M. O., Hansen, R. S., Navas, P. A., Stamatoyannopoulos, G., Iyer, V. R., Lieb, J. D., Sunyaev, S. R., Akey, J. M., Sabo, P. J., Kaul, R., Furey, T. S., Dekker, J., Crawford, G. E. and Stamatoyannopoulos, J. A. (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75–82.

Tolhuis, B., Palstra, R.-J., Splinter, E., Grosveld, F. and de Laat, W. (2002). Looping and Interaction between Hypersensitive Sites in the Active β-globin Locus. Mol. Cell *10*, 1453–1465.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S. and Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat. Biotechnol. *32*, 381–386.

Tseng, W. C., Haselton, F. R. and Giorgio, T. D. (1997). Transfection by cationic liposomes using simultaneous single cell measurements of plasmid delivery and transgene expression. J. Biol. Chem. *272*, 25641–25647.

Väremo, L., Nielsen, J. and Nookaew, I. (2013). Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. Nucleic Acids Res. *41*, 4378–4391.

Vierstra, J., Rynes, E., Sandstrom, R., Zhang, M., Canfield, T., Hansen, R. S., Stehling-Sun, S., Sabo, P. J., Byron, R., Humbert, R., Thurman, R. E., Johnson, A. K., Vong, S., Lee, K., Bates, D., Neri, F., Diegel, M., Giste, E., Haugen, E., Dunn, D., Wilken, M. S., Josefowicz, S., Samstein, R., Chang, K.-H., Eichler, E. E., De Bruijn, M., Reh, T. A., Skoultchi, A., Rudensky, A., Orkin, S. H., Papayannopoulou, T., Treuting, P. M., Selleri, L., Kaul, R., Groudine, M., Bender, M. A. and Stamatoyannopoulos, J. A. (2014). Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. Science *346*, 1007–1012.

Visel, A., Minovitsky, S., Dubchak, I. and Pennacchio, L. A. (2007). VISTA Enhancer Browser—a database of tissue-specific human enhancers. Nucleic Acids Res. *35*, D88–D92.

Vitak, S. A., Torkenczy, K. A., Rosenkrantz, J. L., Fields, A. J., Christiansen, L., Wong, M. H., Carbone, L., Steemers, F. J. and Adey, A. (2017). Sequencing thousands of single-cell genomes with combinatorial indexing. Nat. Methods *14*, 302–308.

Vogelstein, B. and Kinzler, K. W. (1999). Digital PCR. Proceedings of the National Academy of Sciences *96*, 9236–9241.

Wei, K., Iyer, R. and Bilmes, J. (2015). Submodularity in Data Subset Selection and Active Learning. In International Conference on Machine Learning pp. 1954–1963,.

Xie, S., Cooley, A., Armendariz, D., Zhou, P. and Hon, G. C. (2018). Frequent sgRNA-barcode recombination in single-cell perturbation assays. PLoS One *13*, e0198635.

Xie, S., Duan, J., Li, B., Zhou, P. and Hon, G. C. (2017). Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells. Mol. Cell *66*, 285–299.e5.

Yahi, A., Lappalainen, T., Mohammadi, P. and Tatonetti, N. (2018). RecNW: A fast pairwise aligner for targeted sequencing. bioRxiv , 371989.

Yang, M., Büsche, G., Ganser, A. and Li, Z. (2013). Morphology and quantitative composition of hematopoietic cells in murine bone marrow and spleen of healthy subjects. Ann. Hematol. *92*, 587–594.

Yeganeh, M., Praz, V., Cousin, P. and Hernandez, N. (2017). Transcriptional interference by RNA polymerase III affects expression of the Polr3e gene. Genes Dev. *31*, 413–421.

Yu, H., Jetzt, A. E., Ron, Y., Preston, B. D. and Dougherty, J. P. (1998). The nature of human immunodeficiency virus type 1 strand transfers. J. Biol. Chem. *273*, 28384–28391.

Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., Sandstrom, R., Ma, Z., Davis, C., Pope, B. D., Shen, Y., Pervouchine, D. D., Djebali, S., Thurman, R. E., Kaul, R., Rynes, E., Kirilusha, A., Marinov, G. K., Williams, B. A., Trout, D., Amrhein, H., Fisher-Aylor, K., Antoshechkin, I., DeSalvo, G., See, L.-H., Fastuca, M., Drenkow, J., Zaleski, C., Dobin, A., Prieto, P., Lagarde, J., Bussotti, G., Tanzer, A., Denas, O., Li, K., Bender, M. A., Zhang, M., Byron, R., Groudine, M. T., McCleary, D., Pham, L., Ye, Z., Kuan, S., Edsall, L., Wu, Y.-C., Rasmussen, M. D., Bansal, M. S., Kellis, M., Keller, C. A., Morrissey, C. S., Mishra, T., Jain, D., Dogan, N., Harris, R. S., Cayting, P., Kawli, T., Boyle, A. P., Euskirchen, G., Kundaje, A., Lin, S., Lin, Y., Jansen, C., Malladi, V. S., Cline, M. S., Erickson, D. T., Kirkup, V. M., Learned, K., Sloan, C. A., Rosenbloom, K. R., Lacerda de Sousa, B., Beal, K., Pignatelli, M., Flicek, P., Lian, J., Kahveci, T., Lee, D., Kent, W. J., Ramalho Santos, M., Herrero, J., Notredame, C., Johnson, A., Vong, S., Lee, K., Bates, D., Neri, F., Diegel, M., Canfield, T., Sabo, P. J., Wilken, M. S., Reh, T. A., Giste, E., Shafer, A., Kutyavin, T., Haugen, E., Dunn, D., Reynolds, A. P., Neph, S., Humbert, R., Hansen, R. S., De Bruijn, M., Selleri, L., Rudensky, A., Josefowicz, S., Samstein, R., Eichler, E. E., Orkin, S. H., Levasseur, D., Papayannopoulou, T., Chang, K.-H., Skoultchi, A., Gosh, S., Disteche, C., Treuting, P., Wang, Y., Weiss, M. J., Blobel, G. A., Cao, X., Zhong, S., Wang, T., Good, P. J., Lowdon, R. F., Adams, L. B., Zhou, X.-Q., Pazin, M. J., Feingold, E. A., Wold, B., Taylor, J., Mortazavi, A., Weissman, S. M., Stamatoyannopoulos, J. A., Snyder, M. P., Guigo, R., Gingeras, T. R., Gilbert, D. M., Hardison, R. C., Beer, M. A., Ren, B. and Mouse ENCODE Consortium (2014). A comparative encyclopedia of DNA elements in the mouse genome. Nature *515*, 355–364.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W. and Liu, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., Bharadwaj, R., Wong, A., Ness, K. D., Beppu, L. W., Deeg, H. J., McFarland, C., Loeb, K. R., Valente, W. J., Ericson, N. G., Stevens, E. A., Radich, J. P., Mikkelsen, T. S., Hindson, B. J. and Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. Nat. Commun. *8*, 14049.

Zhou, B., Ho, S. S., Greer, S. U., Zhu, X., Bell, J. M., Arthur, J. G. and others (2018). Comprehensive, integrated, and phased whole-genome analysis of the primary ENCODE cell line K562. BioRxiv .

VITA

I grew up in Salem, Connecticut and graduated from East Lyme High School, a public high school in East Lyme, Connecticut. My parents were scientists in the pharmaceutical industry, which gave me the privilege of an early exposure to biology and science in general, although I never thought I would do a PhD.

As a kid, I mostly did well in school, but was never close to the top of my class. I wasn't (and am still not) the most talkative person, but have always had a lot of physical energy, which probably put an above-average burden on my elementary and middle school teachers. To mitigate this, my parents let me play lots of sports and I eventually settled on gymnastics. My family was kind enough to take me to 3-4 hour practices five times a week at Tritown gymnastics in Tolland, CT – a 45 minute drive from our home. Over the years (age 6 to 21), I managed to become quite good and traveled (again, because of the great kindness my family showed me) across the country for competitions, even to national-level ones on several occasions. Having been a gymnast has served as an immensely positive force for good in my life in many ways despite having made a lot of compromises along the way to maintain it as such a large part of my (and my family's) life.

I didn't have much time to study or do anything much beyond my homework assignments, and even that was a stretch at times. All through school, I felt like I was largely trying to keep my head above water and honestly don't recall thinking nearly as much about what I would do academically as what I would do athletically. I loved technology and took pretty much every class I could manage at my high school that sounded vaguely related to engineering and digital arts, but I didn't have time for much else.

My first year of college at the University of Washington was much the same, with gymnastics taking up the majority of my time (25-30 hours a week) and energy. Despite my love for the sport, I decided to stop during my Sophomore year for a number of reasons and it suddenly felt like I had all the time in the world. While I still did sports, I took on a much stronger interest in my classes and actually had time to study. I even joined a robotics lab headed by Blake Hannaford and Howard Chizeck and worked 10-15 hours a week. Every year I did better and better in school and in the lab, although it still seemed impossible to me that I could achieve even a fraction that which many students around me seemed capable. I ended up graduating with a degree in bioengineering, which gave me broad exposure to many different fields of biology and engineering. Funnily enough, my worst grade in all of college was molecular biology, so I was pretty sure I should just become an engineer.

During my time as an undergrad I was fortunate enough to meet my life partner, Molly. Molly had been at the top of her class in high school and was beloved by the Biology department at UW, but you would never know it from the way she acted. Molly was one of the first people outside my family who truly encouraged me academically and personally. She made me start to believe I could be more than just mildly successful outside of athletics.

After graduating, I worked at a medical device company in Boston, but I was bored and unhappy after six months or so. Everyone around me was an engineer and knew pretty much no biology, which wasn't at all satisfying to me (at this point I had all but forgotten about my molecular biology class), and the technologies they were working on felt out of date and unexciting. Molly was working in a genetics lab at Boston Children's Hospital and forced me to talk to the husband of one of her co-workers at a birthday party, who did something called bioinformatics, even though I insisted that I wasn't ready. After our conversation, I ended up volunteering to do a project on the side and enjoyed it so much that I eventually quit my engineering job to join his lab for pretty minimal pay.

The "husband" above turned out to be Monkol Lek, who led the ExAC project in Daniel MacArthur's lab at MGH and the Broad Institute, although none of those details meant much to me at the time. I was lucky enough to have him as my first mentor in genomics and effectively got paid (even if not that much) to learn on the job for the next year. Molly and I both eventually decided that we wanted to go to graduate school and did visits all around the country. In retrospect I was totally unprepared for interviews, only having joined the MacArthur lab in earnest just a few months beforehand, but thankfully Daniel was, and continues to be, an incredible advocate of mine. I was, however, extremely excited and learned enough over the next year to achieve a pretty good level of competency. I don't think I can ever thank Monkol and Daniel enough. There is really no good reason they should have taken me over anyone else, other than my association with Molly, and I still regularly reflect on the arbitrary nature of it all. The progress I made in their group both with respect to skills and self-confidence led me to the first point in my life where I really felt like I could contribute meaningfully within a group of such smart and passionate people.

Molly and I decided to return to Seattle and pursue a PhD in genomics. This was a really hard decision for Molly and I, but Genome Sciences felt like the right combination of excellent science and kind people. Looking back on it now, I can very confidently say that it was the right choice. I'm not sure I'll ever get used to the fact that so many people like Monkol and Daniel, Jay and Cole, all my teammates at 10x genomics, and numerous others continue to advocate for me and endorse my work. While I realize I've come a long way in a short period of time, and am likely deserving of some portion of the attention they've given me, I am still flattered and immensely appreciative of all the kindness that my colleagues have shown me and am very lucky that things have gone my way so many times over.