Epigenomic profiling of human tissues at single-cell resolution

Steven Wu

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

University of Washington
2022

Reading Committee:
Steven Henikoff, Chair
Georg Seelig
William Noble
Manu Setty

Program Authorized to Offer Degree:
Molecular Engineering & Sciences

©Copyright 2022
Steven Wu

University of Washington

Abstract

Epigenomic profiling of human tissues at single-cell resolution

Steven Wu

Chair of the Supervisory Committee:
Steven Henikoff
Affiliate Professor, Department of Genome Sciences

Traditional methods for profiling DNA-protein binding interactions have been limited by low signal-to-noise, false positives, and high costs. To overcome these barriers we developed a simple assay, Cleavage Under Targets & Tagmentation (CUT&Tag), that leverages a transposon based fusion enzyme to map in situ DNA-protein interactions in small samples of cells at high resolution. We then automated CUT&Tag to generate hundreds of chromatin profiles for a multitude of histone modifications across different diseases. Furthermore, we are able to model their cell-type specific gene expression by integrating the data across multiple histone modifications.

CUT&Tag is characterized by an exceptionally high signal-to-noise ratio, and we reasoned that the method could be used to resolve single-cell chromatin profiles. As a proof-of-concept we demonstrated that single-cell CUT&Tag resolves both active and repressive chromatin marks in cell lines. We then leveraged single-cell CUT&Tag to profile thousands of single cells to uncover the heterogeneity in stem cell development, primary liquid, and solid tumors pre- and post-treatment. Our work is part of a large-scale effort to build a comprehensive map of all cell types to better understand human health and improve disease diagnosis and treatment.

Acknowledgements

Thank you, Steve & Kami and the rest of the Henikoff lab iteration from 2018-2022 for making graduate school fun! I would also like to thank Anoop Patel, Scott Furlan, and all my other manuscript co-authors. None of this would be possible without everyone's support.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

Introduction

Chromatin & Gene Regulation

Nucleosomes are the basic repetitive subunits of chromatin (Luger et al., 2012). Composed of 147bp of DNA wrapped around a histone octamer, they are highly dynamic and serve as a hub for gene regulation (Lai & Pugh, 2017). At the most fundamental level of gene regulation, nucleosomes impede transcription machinery such as polymerase II from spuriously binding to the DNA and initiating transcription along any random sequence (Lai & Pugh, 2017). Furthermore, nucleosomes act as a substrate for additional levels of regulation (Bannister & Kouzarides, 2011). An example of this are histone modifications along the lysine residue on the histone 3 (H3) or histone 4 (H4) tail.

Histone Modifications

Diverse sets of histone modifications increase the regulatory complexity of gene regulation (Bannister & Kouzarides, 2011). One example is H3-lysine27 acetylation (H3K27ac). H3K27ac is associated with euchromatin, a gene rich and transcriptionally active site in the genome (Creyghton et al., 2010). Alternatively, trimethylation of H3-lysine 9/27 (H3K9me3/H3K27me3) is associated with heterochromatin; ie. condensed chromatin that is transcriptionally repressed (Cedar & Bergman, 2009). Although these histone modifications have been shown to be correlative with various transcriptional states, the question of causation or consequence and functionality remains quite controversial (Creyghton et al., 2010). One argument for functionality is that the acetyl group is negatively charged like DNA (Bannister & Kouzarides, 2011). The charge repulsions between the acetyl and DNA causes local relaxation which makes room for transcription factors to bind and initiate transcription (Bannister & Kouzarides, 2011). On the other hand, the N-tail terminal knockout of H3 and H4 in yeast yielded no transcriptional consequence (Ling et al., 1996). Despite these contested points about function, histone modifications are an easy way to read out the transcriptional state of chromatin. This is especially true with the rapidly expanding repertoire of chromatin profiling assays (Skene et al., 2018).

Chromatin Profiling

In the early 21st century microarray based approaches were the standard for chromatin profiling (Pollack et al., 1999). Chromatin immunoprecipitation followed by genomic tile microarray hybridization (ChIP-chip) was one of the first high throughput methods to generate high-resolution, genome-wide maps of protein-DNA interactions in smaller genomes, such as yeast (Ren et al., 2000). However, with the completion of the human genome sequencing project and the rise of next-generation sequencing (NGS), the genomics field quickly shifted from microarrays to sequencing based approaches (International

Human Genome Sequencing Consortium et al., 2001). Interestingly, the cost of NGS rapidly decreased following a similar trend to Moore's law (Morey et al., 2013). The decrease in cost made it accessible to most scientists at any institution, which led to a suite of sequencing based assays for measuring the different modalities of chromatin.

A second generation ChIP assay, ChIP followed by sequencing (ChIP-seq), quickly became the gold standard for profiling DNA-protein complexes due to higher resolution, signal-to-noise, dynamic range, and genomic coverage relative to ChIP-chip (P. J. Park, 2009). To date there have been over 100,000 papers published referencing ChIP-seq. There are also high-throughput sequencing based assays for reading de novo genomes (WGS), nucleosome phasing and occupancy (MNase-seq), DNA methylation (WGBS), DNA accessibility (DNase-seq, ATAC-seq), and many more (Reuter et al., 2015). Chromatin profiling is an umbrella term that can refer to any of these assays, but from here on out it will be used to specifically refer to assays that profile DNA-protein interactions.

ChIP-seq has been the dominant assay for probing protein-DNA interactions due to its inherent flexibility -- in theory, the ChIP-seq workflow should be able to map any protein that binds to DNA in a robust manner (P. J. Park, 2009). However, the inherent downside is that only proteins with high-quality antibodies can be mapped and epitope masking from cross-linking generates artifacts (Kidder et al., 2011). In addition, the workflow of ChIP-seq requires high cell numbers, high sequencing depth, and takes several laborious days to conduct each assay (Skene et al., 2018).

Advances in orthogonal methods that leverage fusion enzymes has since rendered ChIP-seq an outdated method (Skene & Henikoff, 2017). DamID (DNA adenine methyltransferase) fused to a chromatin protein was one of first fusion enzymes used to map DNA-protein binding (Steensel & Henikoff, 2000). The chromatin targeted protein localizes to its binding site, bringing Dam in the process and methylating adenine in the sequence GATC. Protein-DNA binding sites can then be inferred from where methylation occurs. However, the spatial resolution of DamID is limited by the GATC sequence frequency, which is every ~200-300bp on average (Steensel & Henikoff, 2000).

The next generation of fusion enzymes were based on micrococcal nuclease (MN), an exogenous endo-nuclease that digests protein-unbound DNA (Keene & Elgin, 1981). One example is chromatin immunocleavage (ChIC), a new assay that utilizes the fusion enzyme (pA-MNase) where protein A (pA) a protein with high affinity for immunoglobulins is tethered to micrococcal nuclease (MN) (Schmid et al., 2004b). In the workflow for ChIC, cells are permeabilized, primary antibody corresponding to the protein of interest is flowed in followed by the fusion enzyme, and then MN is activated by the addition of divalent cations (Ca2+) (Schmid et al., 2004a). Active MNase then cleaves around the surrounding protein of interest. An improved iteration of the ChIC workflow with a sequencing adaption was developed a decade later and dubbed Cleavage Under Target & Release Using Nuclease (CUT&RUN) (Skene & Henikoff, 2017).

Single-Cell Sequencing

The first proof-of-concept experiment for sequencing mRNA in single-cells was published in 2009 (Tang et al., 2009). Prior, a commonality that united all genomic sequencing assays was that the readout of these assays represented a bulk population (Trapnell, 2015). This means that the readout from the assay is averaged across all cells thus losing the ability to describe subpopulations in the sample. This is an inherent flaw with bulk assays because biological samples are generally heterogenous. One example of heterogeneity is in tumors (Patel, 2014). They exhibit a complex environment with a multitude of different subcellular populations such as immune response cells, differentiated tumor cells, and malignant proliferating cancer cells (Patel, 2014). Even in the cases of homogenous samples, published works have shown that individual cells can exhibit significant variation between their protein levels, phenotype, and genotype (Altschuler & Wu, 2010).

Single-cell assays are extremely difficult to perform due to the technical constraints of isolating and performing an assay on a single cell. The first single-cell sequencing experiment profiled the transcriptome of a single mouse blastomere cell by isolating the cell under a microscope and subsequently running the mRNA-seq assay (Tang et al., 2009). In the following years, technological advances in mRNA assays, notably smart-seq, increased the readout efficiency of full length mRNA transcripts from small input (Picelli et al., 2014). Soon after, a multitude of single cell RNA-seq (scRNA-seq) manuscripts appeared as labs were able to robustly profile hundreds of single-cells for different biological systems (Patel, 2014; Trapnell et al., 2014). In that same year, Nature Methods voted single-cell sequencing assays as the method of the year (*Method of the Year 2013* | *Nature Methods*, n.d.).

Despite major advancements in scRNA-seq, the throughput or number of single-cell recovered from each assay was low and at best in the ~100s of cells. This meant that to capture 100 single-cell profiles, 100 separate RNA-seq assays were performed in parallel which is quite laborious and prone to batch effects (A. R. Wu et al., 2014). Technological advances made it possible to automate hundreds of these single-cell RNA-seq assays in parallel and reduce batch effect. One notable instrument for this is Takara's iCell8 nanowell platform (Gao et al., 2017). The iCell8 dispenses hundreds to thousands of cells into a microchip containing 5184 nanowells using a poisson distribution to approximate the concentration of cells that equates to one cell per nanowell. Subsequently, imaging is conducted to include wells with only one single-cell, then barcode for each single-cell well and subsequently PCR for library construction (Gao et al., 2017).

The iCell8 automated hundreds to a thousand single-cell reactions in parallel with low batch effects. However, it wasn't until the advent of microfluidics, namely drop-seq, when the number of single-cells that could be sequenced scaled from the low thousands to the tens of thousands (Macosko et al., 2015). Drop-Seq encapsulates single-cells in oil droplets with microparticles that are conjugated to unique barcodes for each particle, cells then are lysed within the oil droplet, hybridization of the unique barcode with RNA from the lysed cell occurs, followed by breaking the oil droplet and reverse transcription in bulk (Macosko et al., 2015).

The latest advancements in single-cell genomics came from combinatorial indexing, which increased the throughput of single-cell assays from ~thousands to millions (Cusanovich et al., 2015). In combinatorial indexing, cells are split into different wells with unique primers in each well. Cells in the same well all have their transcripts labeled with the same primer via PCR. The cells are then pooled and redistributed at random to different wells for another round of barcoding via ligation and PCR. Unique single-cell identifiers are generated after N rounds of barcoding by loading the wells an order of magnitude less than possible combinations (Cusanovich et al., 2015; Rosenberg et al., 2018). Combinatorial indexing has enabled scientists to generate transcriptome atlases of both mouse and human fetal development for millions of single-cells (Cao et al., 2019, 2020).

The technological breakthroughs for scRNA-seq were applied to other sequencing methods in the following years. The most notable example is Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) (Buenrostro et al., 2013). ATAC-seq uses a hyperactive Tn5 Transposase that inserts sequencing adapters into DNA accesibile sites. By 2015, single-cell ATAC-seq had been adapted to both microfluidics platforms and combinatorial indexing (Buenrostro, 2015; Cusanovich et al., 2015). Soon thereafter, chromatin accesiblity atlasing projects for hematopoiesis, human fetal development, and various cancers were published (Corces et al., 2016; Domcke et al., 2020).

ATAC-seq is a simple method to catalog active regulatory elements and was the first single-cell method to profile any modality of chromatin (Buenrostro, 2015; Cusanovich et al., 2015). However, single-cell assays to profile specific protein-DNA interactions lagged behind in development. One reason for this could be technical complexity of a standard DNA-protein interaction assay, i.e ChIP-seq, vs. ATAC-seq. Furthermore, since nuclei have to be lysed for the ChIP process, ChIP-seq can't be adapted to combinatorial indexing. However, advancements in drop-seq, dubbed drop-ChIP, generated the first single-cell profiles for H3K4me3 at active promoters in embryonic stem cells (Rotem, 2015). In the following years protein-DNA methods utilizing fusion enzymes such as CUT&RUN and ChIC-seq were adapted to low through-put methods of single-cell profiling (Hainer et al., 2018; Ku et al., 2019). The workflow for CUT&RUN and ChiC-seq generates free floating immune-targeted protein-DNA fragments that diffuse out of the nuclei into the supernatant. Therefore, these methods of protein-DNA interactions aren't scalable using methods like combinatorial indexing.

Description of Thesis

When I joined the lab, Pete Skene and Steve Henikoff had just published a simple and robust workflow for epigenomic profiling: CUT&RUN (Skene et al., 2018; Skene & Henikoff, 2017). At that time, Jay Sarthy, a M.D/PhD fellow, had just joined the lab and asked how we could adapt basic research tools to a more clinically relevant setting. Simultaneously, academic interest in single-cell sequencing technology was ballooning as well. Needless to say, there was a lot of interest in various scientific directions we could take. Steve described this period as the "one of the most exciting time in the lab". In my thesis, I'll describe how we leverage CUT&RUN and adapt it to a clinical setting, improve the simplicity

and scalability of chromatin profiling, adapt epigenomic profiling to single-cell resolution, and finally, use single-cell resolution to conduct research on clinically relevant samples.

During my rotation, a postdoc named Derek Janssens automated CUT&RUN by adapting the protocol to a 96 well-plate format that could be integrated into a biomek workflow (Janssens, 2018). In chapter 2, I will describe how we take advantage of autoCUT&RUN to devise a simple yet cost efficient metric to model cell type specific gene expression by integrating multiple histone modifications. Using autoCUT&RUN we are able to profile any sample for about ~$100 including sequencing cost. We benchmark our data to “gold standard” ChIP-seq data from the ENCODE project. Furthermore, we show that our data for “active” histone modifications such as H3K4me1-3 and H3K27ac are highly correlated with RNA-seq data. Based on this observation, we were able to model the transcriptional activity using a simple linear regression. The weights from our model reaffirmed our prior observation that some histone modifications are positively associated with transcription, whereas other silencing histone modifications such as H3K27me3 are negatively associated. We then test our model in two different patient derived cell lines and uncover the differences between their subtype-specific regulatory elements.

Next, we benchmarked H3K4me2, a mark of promoters and enhancers, in comparison to ATAC-seq. Despite higher signal-to-noise in CUT&RUN H3K4me2 compared to ATAC-seq, ATAC-seq is used more frequently. This is likely due to the sheer ease of the assay and the ability to run it in single-cell resolution. A postdoc named Hatice Kaya-Okur further simplified the tethering enzyme chromatin method, by using the Tn5 transposase in ATAC-seq and fusing it to protein-A (Kaya-Okur, 2019). This new complex, pA-Tn5, drops sequencing adapters directly into the antibody targeted site, thus removing the need for a library preparation step. In Chapter 3, I describe how this pA-Tn5 fusion enzyme allows for a cost-efficient method, Cleavage Under Targets & Tagmentation (CUT&Tag), in epigenomic profiling. We benchmark CUT&Tag against other methods of chromatin profiling methods and demonstrate a higher signal-to-noise ratio for profling polymerase, histone modifications, and transcription factors.

Single-cell profiling is limited by sparsity, but since our signal-to-noise was so high in CUT&Tag for bulk samples, we rationalized that our method would hold in single-cells as well. Utilizing the iCell8 nano-dispense machine, we were able to profile active and repressive histone modifications in single-cells for K562 and hESC cells and discern their cell-type from their chromatin profiles. In Chapter 4, we profile polycomb silenced regions, marked by H3K27me3, in single-cell resolution to uncover differentiation and tumor progression (S. J. Wu et al., 2021). We profile a 5 day time course of differentiating hESC to definitive endoderm in 24 hour steps. The single-cell H3K27me3 profiles revealed a developmental trajectory, thus showing that our method could profile developmental systems. Next, we profiled peripheral blood mononuclear cells to see if scCUT&Tag had any biases for different celltypes. We showed that there were no biases in our single-cell method and recovered cell type proportions similar to fluorescence-activated cell sorting.

We then demonstrated the utility of single-cell CUT&Tag in clinical samples by profiling a primary human glioblastoma sample prior to treatment and the same relapsed sample (Janssens, Meers, et al., 2021). We uncovered a tumor population of cells whose chromatin profile suggest they are undergoing a pro-neuronal to mesenchymal shift. In Chapter 5, we then apply scCUT&Tag to mixed-lineage leukemias with chromosomal translocation involving lysine methyltransferase (KMT2a). First, we used bulk CUT&Tag to characterize the regulatory landscape of mixed phenotype leukemia and identify direct KMT2a targets. Integrating direct target sites with chromatin profiles for H3K27me3 and H3K4me3 allowed us to identify bivalent sites that were targeted by KMT2a. We then profiled the KMT2a leukemias in single-cell resolution for histone modifications H3K27me3, H3K36me3, and H3K27me3. We then observed how chromatin bivalency is tied to heterogeneity within a sample. In Chapter 6, I will discuss the significance of this work and newly emerging single-cell technology in the “multi-omic” space.

Chapter 2

Automated in situ chromatin profiling efficiently resolves cell types and gene regulatory programs

Modified from an article originally published in *Epigenetics & Chromatin:*
Derek H. Janssens, Steven J. Wu, Jay F. Sarthy, Michael P. Meers, Carrie H. Myers, James M. Olson, Kami Ahmad & Steven Henikoff. Epigenetics & Chromatin volume 11, Article number: 74 (2018)

Abstract

Our understanding of eukaryotic gene regulation is limited by the complexity of protein–DNA interactions that comprise the chromatin landscape and by inefficient methods for characterizing these interactions. We recently introduced CUT&RUN, an antibody-targeted nuclease cleavage method that profiles DNA-binding proteins, histones and chromatin-modifying proteins in situ with exceptional sensitivity and resolution.

Here, we describe an automated CUT&RUN platform and apply it to characterize the chromatin landscapes of human cells. We find that automated CUT&RUN profiles of histone modifications crisply demarcate active and repressed chromatin regions, and we develop a continuous metric to identify cell-type-specific promoter and enhancer activities. We test the ability of automated CUT&RUN to profile frozen tumor samples and find that our method readily distinguishes two pediatric glioma xenografts by their subtype-specific gene expression programs.

The easy, cost-effective workflow makes automated CUT&RUN an attractive tool for high-throughput characterization of cell types and patient samples.

Introduction

Cells establish their distinct identities by altering activity of the cis-regulatory DNA elements that control gene expression (Heinz et al., 2015; Levine et al., 2014). Promoter elements lie near the 5′ transcriptional start sites (TSSs) of all genes, whereas distal cis-regulatory elements such as enhancers often bridge long stretches in the DNA to interact with select promoters and direct cell-type-specific gene expression (Heinz et al., 2015; Levine et al., 2014). Defects in the nuclear proteins that recognize these cis-regulatory elements underlie many human diseases that often manifest in specific tissues and cell types (Cotney et al., 2015; Hu & Shilatifard, 2016; Lambert et al., 2018; Mackay et al., 2017; Schwartzentruber et al., 2012). However, we are only just beginning to appreciate how assessing the activity of cis-regulatory elements may be used in clinical settings for patient diagnosis (Corces et al., 2018). To provide a reference for molecular diagnosis of patient samples, efforts are underway to generate a comprehensive atlas of cells in the human body (Regev et al., 2017; Rozenblatt-Rosen et al., 2017). Characterizing cell-type-specific chromatin landscapes is essential for this atlas; however,

technical limitations have prevented implementation of traditional approaches for genome-wide profiling of chromatin proteins on the scales necessary for this project.

Despite the growing awareness that epigenetic derangements underlie many human diseases (Feinberg, 2018), very few methods for high-throughput profiling of epigenomic information are available. Realizing the clinical potential of epigenomic technologies requires robust, scalable approaches that can profile large numbers of patient samples in parallel. Chromatin immunoprecipitation with antigen-specific antibodies combined with massively parallel sequencing (ChIP-seq) has been used extensively for epigenome profiling, but this method is labor-intensive, prone to artifacts (Jain et al., 2015; D. Park et al., 2013; Teytelman et al., 2013), and requires high sequencing depth to distinguish weak signals from genomic background noise. Although semi-automated implementations of ChIP-seq exist, these begin with cross-linking cells and solubilization by sonication (Aldridge et al., 2013; Gasper et al., 2014; Wallerman et al., 2015), steps that are difficult to scale and to control for reproducibility. The combination of these factors has prevented implementation of ChIP-seq in clinical laboratory settings. Recently, we have introduced CUT&RUN as an alternative chromatin profiling technique that uses factor-specific antibodies to tether micrococcal nuclease (MNase) to genomic binding sites (Skene et al., 2018; Skene & Henikoff, 2017). The targeted nuclease cleaves chromatin around the binding sites, and the released DNA is sequenced using standard library preparation techniques, resulting in efficient mapping of protein-DNA interactions. CUT&RUN has very low backgrounds, which greatly reduces sample amounts and sequencing costs required to obtain high-quality genome-wide profiles (Hainer et al., 2018; Skene et al., 2018).

Here, we modify the CUT&RUN protocol to profile chromatin proteins and modifications in a 96-well format on a liquid handling robot, beginning with permeabilized cells and ending with barcoded libraries that are ready to be pooled for sequencing. By applying this method to the H1 human embryonic stem cell (hESC) line and the K562 leukemia cell line, we demonstrate that AutoCUT&RUN can be used to identify cell-type-specific promoter and enhancer activities, providing a means to quantitatively distinguish cell-types based on their unique gene regulatory programs. In addition, we show that this method is able to define chromatin features from frozen solid tumor samples, setting the stage to analyze typical clinical specimens at low cost. AutoCUT&RUN is ideal for high-throughput studies of chromatin-based gene regulation, allowing for examination of chromatin landscapes in patient samples and expanding the toolbox for epigenetic medicine.

Results

An automated platform for genome-wide profiling of chromatin proteins

To adapt CUT&RUN to an automated format we equipped a Beckman Biomek FX liquid handling robot for magnetic separation and temperature control (Figure 2.1a). First, cells are bound to concanavalin A-coated magnetic beads, allowing all subsequent washes to be performed by magnetic separation. Bead-bound samples are then incubated with antibodies, and up to 96 samples are arrayed in

a plate (Figure 2.1a). Successive washes, tethering of a proteinA-MNase fusion protein, cleavage of DNA, and release of cleaved chromatin fragments into the sample supernatant are performed on the Biomek (Figure 2.2a). A major stumbling block to automating epigenomics protocols is that they typically require purification of small amounts of DNA prior to library preparation. To overcome this hurdle, we developed a method to polish the DNA ends in chromatin fragments for direct ligation of Illumina library adapters (Figure 2.2a). End-polishing and adapter ligation are performed on a separate thermocycler, and deproteinated CUT&RUN libraries are purified on the Biomek using Ampure XP magnetic beads both before and after PCR enrichment. This AutoCUT&RUN protocol allows a single operator to generate up to 96 libraries in 2 days that are ready to be pooled and sequenced (Figure 2.1a) (https://www.protocols.io/view/autocut-run-genome-wide-profiling-of-chromatin-pro-ufeetje).

To test the consistency of AutoCUT&RUN, we simultaneously profiled two biological replicates of H1 hESCs and K562 cells using antibodies targeting four histone modifications that mark active chromatin states (H3K4me1, H3K4me2, H3K4me3, and H3K27ac) and one repressive modification (H3K27me3). Comparing the global distribution of reads for each histone mark, we found that samples highly correlate with their biological replicate and cluster together in an unbiased hierarchical matrix (Figure 2.1b). Additionally, the genome-wide profiles of the four active histone marks clustered together within a given cell type and separated away from the repressive histone mark H3K27me3 (Figure 2.1b). These profiles represent antibody-specific signals, as all five are poorly correlated with an IgG-negative control. Together, these results indicate that AutoCUT&RUN chromatin profiling reproducibly captures the cell-type-specific distributions of histone marks.

In addition to profiling histone modifications, we also examined whether AutoCUT&RUN can be applied to mapping DNA-binding transcription factors. We tested the performance of AutoCUT&RUN with two transcription factors, the histone locus-specific gene regulator NPAT, and the insulator protein CTCF (Narendra et al., 2015; Zhao et al., 2000). AutoCUT&RUN profiles of both NPAT and CTCF are highly specific for their expected targets in both H1 and K562 cells (Figure 2.2b, c). Thus, AutoCUT&RUN is suitable for high-throughput, genome-wide profiling of diverse DNA-binding proteins.

Comparison of AutoCUT&RUN to ChIP-seq

We previously showed that the low backgrounds and high efficiency of CUT&RUN allowed for much lower DNA sequencing read depths than required for conventional ChIP-seq to obtain good feature definition. To determine whether improved performance relative to ChIP-seq extends to AutoCUT&RUN, we first identified the histone modification datasets from the ENCODE project that used the same antibodies and manufacturer catalog numbers as we used. A representative region is shown for direct comparison of tracks between AutoCUT&RUN and ENCODE (Figure 2.3a). In all comparisons, the ENCODE datasets are seen to be much noisier than the CUT&RUN datasets, despite the fact that there were ~ 2 to 3 times as many mapped reads in the ENCODE datasets, pooling all of the reads from the two replicates available in the Gene Expression Omnibus (GEO). The requirement for much deeper

sequencing using ChIP-seq relative to CUT&RUN is illustrated by the effect of downsampling the ENCODE datasets to a number of reads equivalent to the number of CUT&RUN fragments, where in the case of H3K4me1, the feature definition from ChIP-seq becomes dramatically reduced, whereas the same number of mapped CUT&RUN fragments shows clear peaks with much lower background than seen for either the Broad Institute or SYDH ENCODE tracks for all comparisons. We confirmed that the higher data quality of CUT&RUN extends to genome-wide analysis, where heat maps of MACS2 peak calls for CUT&RUN show much better signal-to-noise than heat maps for corresponding ENCODE datasets using ENCODE-generated peak calls (Figure 2.3b).

The fixation, sonication, and immunoprecipitation steps of ChIP-seq have the potential to introduce significant batch-effect variability between experiments (Goh et al., 2017; Leek et al., 2010), often making it difficult to directly compare large ChIP-seq datasets generated by different laboratories. To examine whether AutoCUT&RUN reduces this batch-effect variability, we compared the global distribution of reads for H3K4me1 in K562 cells generated from multiple different AutoCUT&RUN and ENCODE ChIP-seq experiments. This analysis revealed that biological replicates profiled by AutoCUT&RUN in different batches have a similar correlation with biological replicates profiled in parallel, indicating there is very little batch-effect variability between AutoCUT&RUN experiments (Figure 2.3c). Furthermore, the correlation between AutoCUT&RUN samples and the Broad Institute ChIP-seq samples was similar to the correlation of Broad Institute ChIP-seq replicates with each other (Figure 2.3c). In agreement with the visual comparison of tracks (Figure 2.3a), this demonstrates the genome-wide profiles of H3K4me1 generated by AutoCUT&RUN are consistent with results obtained using ChIP-seq. However, the correlation between the Broad Institute ChIP-seq replicates and the SYDH ChIP-seq replicate was far lower than that observed between different AutoCUT&RUN batches, reaffirming the inherent difficulty in reproducing ChIP-seq results (Figure 2.3c). We conclude that by eliminating many of the potential sources of batch-effects associated with ChIP-seq, AutoCUT&RUN significantly improves reproducibility between experiments, which will facilitate the adaptation for clinical applications.

We next examined whether AutoCUT&RUN profiles recapitulate global chromatin features that have been previously ascribed to hESCs using ChIP-seq. To maintain their developmental plasticity, hESCs are thought to have a generally “open,” hyper-acetylated chromatin landscape interspersed with repressed domains of “bivalent” chromatin, marked by overlapping H3K27me3 and H3K4 methylation (Bernstein et al., 2006; Gaspar-Maia et al., 2011; Hawkins et al., 2010; Rada-Iglesias et al., 2011). AutoCUT&RUN recapitulates these features of hESCs; we observed that H1 cells have increased H3K27ac as compared to the lineage-restricted K562 cell line, whereas domains of the repressive histone mark H3K27me3 are rare in H1 cells, but prevalent in K562 cells (Figure 2.4a). We also observed extensive overlap between H3K27me3 and H3K4me2 signals in H1 cells, but not K562 cells (Figure 2.4a, b). Thus, AutoCUT&RUN profiles are consistent with the specialized chromatin features found in hESCs using ChIP-seq.

Post-translational modifications to the H3 histone tail closely correlate with transcriptional activity (Karlic et al., 2010). To determine whether our AutoCUT&RUN profiles of histone modifications are indicative of transcriptional activity, we examined the distribution of the five histone marks around the transcriptional start sites (TSSs) of genes, rank-ordered according to RNA-seq expression data (Figure 2.4c, d) ("An Integrated Encyclopedia of DNA Elements in the Human Genome," 2012). We find the active mark H3K4me3 is the most highly correlated with expression in both cell types ($r = 0.70$ and 0.81 for H1 and K562, respectively), followed by H3K4me2 and H3K27ac (Figure 2.4c, d). The repressive histone mark H3K27me3 is anti-correlated with expression ($r = -0.16$ and -0.53 in H1 and K562, respectively) (Figure 2.4c, d). We conclude AutoCUT&RUN for these histone marks provides a strategy to identify cell-type-specific gene regulatory programs.

Modeling cell-type-specific gene expression from AutoCUT&RUN profiles

To use AutoCUT&RUN data to compare cell types and distinguish their gene regulatory programs, we wanted to develop a continuous metric that incorporates both active and repressive chromatin marks. RNA-seq has been widely used to identify cell-type-specific gene expression programs ("An Integrated Encyclopedia of DNA Elements in the Human Genome," 2012), so we used RNA-seq data as a reference for training a weighted linear regression model that incorporates normalized H3K4me2, H3K27ac, and H3K27me3 read counts to assign promoters a relative activity score. We initially focused our analysis on genes with a single TSS that could be unambiguously assigned RNA-seq values. H3K4me2 was selected over H3K4me3 and H3K4me1 because H3K4me2 is uniquely applicable for modeling the activity of both proximal and distal cis-regulatory elements (see below). When applied to K562 cells, promoter chromatin scores correlate very well with RNA-seq values ($r = 0.83$) (Figure 2.5a), providing a comparable power for predicting gene expression as similar models that used up to 39 histone modifications mapped by ChIP-seq ($r = 0.81$) (Karlic et al., 2010). In addition, our weighted model trained on K562 cells performs well when applied to H1 cells (Figure 2.6a, b), indicating that the linear model and data quality are sufficiently robust to assign promoter scores to diverse cell types.

Next, we examined whether AutoCUT&RUN accurately identifies promoters with cell-type-specific activity. By calling promoter scores that were enriched more than twofold in either H1 or K562 cells, we identified 2168 cell-type-specific genes and approximately 40% of these genes (865) were also differentially enriched between H1 and K562 cells according to RNA-seq (Figure 2.5a-d). However, promoter activity modeling did not capture transcriptional differences for 1149 genes (Figure 2.5d and Figure 2.6c, d), implying that these genes are differentially expressed without changes in the chromatin features included in our model. This differential sensitivity between methods suggests the three histone marks included in our chromatin model may more accurately predict the cell-type-specific expression of

certain classes of genes than others. Indeed, we find the 865 cell-type-specific genes identified by both promoter activity modeling and RNA-seq are highly enriched for developmental regulators, whereas the genes called by either promoter scores or RNA-seq alone are not nearly as enriched for developmental GO terms (Figure 2.5d and Figure 2.6e-g). In addition, only 35 genes display contradictory cell-type specificities according to promoter chromatin scores and RNA-seq (Figure 2.5d). This demonstrates AutoCUT&RUN profiling of these widely studied modifications to the H3 histone tail can be applied to accurately distinguish between cell-type-specific developmental regulators.

To determine whether AutoCUT&RUN data recapitulate the expression of cell-type-specific transcription factors, we expanded our analysis to include all promoters. We find that components of the hESC pluripotency network (NANOG, SOX2, SALL4, and OTX2) have higher promoter chromatin scores in H1 cells, while regulators of hematopoietic progenitor cell fate (PU.1, TAL1, GATA1, and GATA2) are enriched in K562 cells (Figure 2.5e) (Gottgens, 2015; Martello & Smith, 2014). This method also identifies activities of alternative promoters (e.g., at the OTX2 and TAL1 genes), providing an indication of the specific gene isoforms that are expressed in a given cell type (Figure 2.5e). We conclude that AutoCUT&RUN can distinguish between master regulators of cellular identity, providing a powerful tool to characterize cell-types in a high-throughput format.

Profiling tumors by AutoCUT&RUN

Typical clinical samples often contain small amounts of material and have been flash-frozen, and although ChIP-seq has been applied to flash-frozen tissue samples, available methods are not sufficiently robust for diagnostic application. In addition, translational samples from xenografts, which are increasingly being used in clinical settings to probe treatment strategies for patients with high-risk malignancies (Malaney et al., 2014). These specimens can be extremely challenging to profile by ChIP-seq as they often contain a significant proportion of mouse tissue and so require extremely deep sequencing to distinguish signal from noise. To test whether AutoCUT&RUN is suitable for profiling frozen tumor specimens, we obtained two diffuse midline glioma (DMG) patient-derived cell lines (VUMC-10 and SU-DIPG-XIII) that were autopsied from similar regions of the brainstem, but differ in their oncogenic backgrounds (Nagaraja et al., 2017). SU-DIPG-XIII is derived from a tumor containing an H3.3K27M “oncohistone” mutation, which results in pathologically low levels of PRC2 activity, and because of this has been called an “epigenetic” malignancy. In contrast, VUMC-10 is a MYCN-amplified, histone wild-type brainstem glioma (Malaney et al., 2014). Both of these DMG cell lines readily form xenografts in murine models, and we applied AutoCUT&RUN to profile histone modifications in VUMC-10 and SU-DIPG-XIII xenografts that were seeded in the brains of mice and then resected upon tumor formation and frozen under typical clinical conditions (Figure 2.7a). For comparison, on the same AutoCUT&RUN plate we profiled the parental DMG cell lines grown in culture (Figure 2.7a). Again, we found that replicates were highly concordant, so we combined them for further analysis. Importantly, cell culture samples were highly correlated with the same mark profiled in the corresponding frozen xenografts, and AutoCUT&RUN on

xenograft tissues and cell culture samples produced similar data quality (Figures 2.7b and Figure 2.8). Thus, AutoCUT&RUN reliably generates genome-wide chromatin profiles from frozen tissue samples.

Stratification of patient malignancies is becoming increasingly dependent on molecular diagnostic methods that distinguish tumor subtypes derived from the same tissues. Our VUMC-10 and SU-DIPG-XIII samples provide an excellent opportunity to explore the potential of using AutoCUT&RUN to classify tumor specimens according to their subtype-specific regulatory elements. By applying promoter modeling to these samples, we identified 5006 promoters that show differential activity between VUMC-10 and SU-DIPG-XIII cells (Figure 2.9a). Consistent with the glial origins of these tumors, both the VUMC-10- and SU-DIPG-XIII-specific promoters are significantly enriched for genes involved in nervous system development (Figure 2.10a, b). Genes involved in cell signaling are also overrepresented in SU-DIPG-XIII cells (Figure 2.10b); for example, the promoters of the PDGFR gene as well as its ligand PDGF are highly active in SU-DIPG-XIII cells (Figure 2.9a). This is consistent with the observation that DMGs frequently contain activating mutations in PDGFR-α that promote tumor growth (Mackay et al., 2017). In addition, one promoter of the SMAD3 gene, a component of the TGF-β signaling pathway (Massague & Chen, 2000), is specifically active in SU-DIPG-XIII cells, whereas two different SMAD3 promoters are active in VUMC-10 cells (Figure 2.8 and Figure 2.9a). In comparison, our model indicates that only 388 promoters differ between VUMC-10 xenografts and cultured cells, and 1619 promoters differ between SU-DIPG-XIII samples (Figure 2.9b,Figure 2.11c). In addition, comparing promoter chromatin scores in an unbiased correlation matrix also indicates DMG xenografts are far more similar to their corresponding cell culture samples than they are to other DMG subtypes or to H1 or K562 cells (Figure 2.9c). This suggests that AutoCUT&RUN can be applied to identify promoters that display tumor subtype-specific activity, providing a reliable method to assign cellular identities to frozen tumor samples, as well as an indication of the signaling pathways that may be driving tumor growth and potential susceptibility to therapeutic agents.

High-throughput mapping of cell-type-specific enhancers

The cell-type-specific activities of gene promoters are often established by incorporating signals from distal cis-regulatory elements, such as enhancers (Heinz et al., 2015; Levine et al., 2014). Similar to promoters, enhancers also display H3K4me2 (Heintzman et al., 2007), and active enhancers are typically marked by H3K27ac, whereas repressed enhancers are marked by H3K27me3 (Bernstein et al., 2006; Creyghton et al., 2010; Heintzman et al., 2009). Therefore, we reasoned that the AutoCUT&RUN profiles we used to model promoter activity should also allow identification of cell-type-specific enhancers. To investigate this possibility, we first compared our H1 data to available chromatin accessibility maps generated by ATAC-seq, which are enriched for both active promoters and enhancers (R. Andersson et al., 2014; Q. Liu et al., 2017). Of the marks we profiled, we find H3K4me2 peaks show the highest overlap with ATAC-seq (Figure 2.11a and Figure 2.12a), and identify 36,725/52,270 ATAC-seq peaks (~ 70%). Interestingly, H3K4me2 defines an additional 71,397 peaks that were not called by ATAC-seq (Figure

2.11a and Figure 2.12a). Many of these H3K4me2-specific peaks show a low, but detectable ATAC-seq signal (Figure 2.11b), indicating they may correspond to repressed promoters and enhancers. Consistent with this interpretation, on average H3K4me2+/ATAC-TSSs have higher H3K27me3 signals than H3K4me2+/ATAC+ TSSs (Figure 2.11c). H3K4me2+/ATAC+ peaks that overlap with annotated TSSs are enriched for H3K4me3, while those peaks that do not overlap TSSs are enriched for H3K4me1 (Figure 2.11d and Figure 2.12b, c), suggesting that many of these distal peaks are enhancers (Calo & Wysocka, 2013; Rada-Iglesias et al., 2011). Thus, mapping sites of H3K4me2 by AutoCUT&RUN provides a sensitive method for defining the repertoire of active cis-regulatory elements that control gene expression programs.

Finally, we examined whether AutoCUT&RUN can be used to identify cell-type-specific enhancers. To expand the number of putative enhancer sites, we compiled a list of non-TSS peaks called on H3K4me2 profiles from all six cell lines and xenograft samples. Using our linear regression model, we then assigned these elements chromatin scores and examined their correlations between different cell types. We find that the chromatin scores of DMG cell culture samples and xenografts are highly correlated ($r = 0.75$ and 0.87 for SU-DIPG-XIII and VUMC-10 samples, respectively) (Figure 2.12d). In contrast, the chromatin scores of SU-DIPG-XIII cells show a weak positive correlation with VUMC-10 cells (e.g. $r = 0.19$), indicating tumor subtype-specific differences. For example, different enhancers near the SOX2 pluripotency gene are active in VUMC-10 cells than SU-DIPG-XIII or H1 cells (Figure 2.12e), indicating that SU-DIPG-XIII cells resemble a more primitive neural stem cell type than VUMC-10 cells, as has been previously suggested (Filbin et al., 2018). Thus, modeling enhancer activity from AutoCUT&RUN profiles of chromatin marks is a highly discriminative method for stratifying cell types and tissue samples to inform patient diagnosis.

Discussion

We adapted the CUT&RUN technique to an automated platform by developing direct ligation of chromatin fragments for Illumina library preparation, and implementing magnetic separation for the wash steps and library purification. AutoCUT&RUN generates 96 genome-wide profiles of antibody-targeted chromatin proteins in just 2 days, dramatically increasing the throughput and potential scale of studies to interrogate the chromatin landscape at a fraction of the cost of comparable lower-throughput methods.

A looming issue in the field of genomics is that extracting meaningful biological insights and clinical information from large datasets is often confounded by batch-effect variability that can arise from numerous sources including different experimental times, reagents, and operators (Goh et al., 2017; Leek et al., 2010). Automated versions of ChIP-seq have been described in which cross-linked and sonicated chromatin is immunoprecipitated on beads for automated library preparation, yielding results that are comparable to manual versions of the same or similar protocol (Aldridge et al., 2013; Gasper et al., 2014; Wallerman et al., 2015). However, cross-linking and sonication are the most difficult steps of a ChIP-seq pipeline to control, and so the non-automated steps of automated ChIP-seq represent a barrier to routine

clinical application, where reproducibility is paramount. Moreover, the stark differences in quality between two different laboratories following the same ENCODE protocols using the same H3K4me1 antibody (Figure 2.3) and subject to the extremely high standards imposed on the ENCODE consortium (Landt et al., 2012) illustrate how difficult it is to obtain uniform data quality in high-throughput ChIP-seq operations. In contrast, AutoCUT&RUN automates the entire process beginning with permeabilized cells or triturated tissues, and returns consistent data that have much better feature definition than that produced by ChIP-seq.

The low backgrounds and high efficiency inherent to antibody-targeted in situ profiling greatly reduce sequencing costs relative to ChIP-seq, surmounting the second major barrier to adoption of genome-wide epigenomic profiling for clinical applications. For example, we estimate that the cost per sample of the datasets we generated for this project was ~ $75 and required 2 days of technician time for 96 samples, ~ 1/10th the cost of commercial whole-exome sequencing (e.g., https://www.abmgood.com/Exome-Sequencing-Service.html). We expect that implementation as a routine service will allow institutional facilities to integrate AutoCUT&RUN into their sequencing pipelines for users who would provide only the cells or tissues and antibodies.

Using CUT&RUN, we have shown that profiling just three histone modifications (H3K27ac, H3K27me3, and H3K4me2) is sufficient to determine the cell-type-specific activities of developmentally regulated promoters and enhancers, providing a powerful quantitative metric to compare the epigenetic regulation of different cell types. This summary metric of chromatin features could be used to assess new cell types and tissue samples and to place them within a reference map of both healthy and diseased cell types. The automated workflow reduces technical and batch-to-batch variability between experiments, generating consistent profiles from biological replicates and from different sample types.

To continue optimizing AutoCUT&RUN, one could envision hardware modifications and computational development. By screening various antibody collections, the repertoire of nuclear proteins that can be efficiently profiled using AutoCUT&RUN would expand dramatically. In addition, the current AutoCUT&RUN protocol is optimized for a popular liquid handling robot, but a custom robot incorporating a reversibly magnetic thermocycler block would allow the CUT&RUN reaction and library preparation to be carried out in place, streamlining the protocol even further. Finally, metrics distinguishing cell types could be improved by incorporating additional aspects of the data, such as using a combination of both enhancer and promoter activities.

The excellent reproducibility of profiling frozen tissue samples by AutoCUT&RUN has the potential to transform the field of epigenomic medicine (Feinberg, 2018). Compared to other genomics approaches that are currently used for patient diagnosis, AutoCUT&RUN has the unique capacity to efficiently profile pathological chromatin proteins within diseased cells. For example, cancers caused by oncogenic chromatin-associated fusion proteins could be profiled by AutoCUT&RUN to provide a molecular diagnosis based on their chromatin landscapes, while simultaneously mapping the loci that are disrupted by the mutant fusion protein. This could provide a powerful tool for patient stratification, as well

as a direct read-out of whether chromatin-modulating therapies such as histone deacetylase or histone methyltransferase inhibitors are having their intended effects.

Methods

AutoCUT&RUN

In conjunction with this work, a detailed AutoCUT&RUN protocol has been made publicly available on Protocols.io (https://www.protocols.io/view/autocut-run-genome-wide-profiling-of-chromatin-pro-ufeetje). Briefly, cells or tissue samples are bound to concanavalin A-coated magnetic beads (Bangs Laboratories, ca. no. BP531), permeabilized with digitonin, and bound with a protein specific antibody as previously described (Skene et al., 2018). Samples are then arrayed in a 96-well plate and processed on a Beckman Biomek FX liquid handling robot equipped with a 96S Super Magnet Plate (Alpaqua SKU A001322) for magnetic separation of samples during wash steps, and an Aluminum Heat Block Insert for PCR Plates (V&P Scientific, Inc. VP741I6A) routed to a cooling unit to perform the MNase digestion reaction at 0–4 °C after the addition of 2 mM CaCl2. MNase digestion reactions are then stopped after 9 min by adding EGTA, which allows Mg2+ addition for subsequent enzymatic reactions. This step is critical for automation because it circumvents the need for DNA purification prior to library preparation. Chromatin fragments released into the supernatant during digestion are then used as the substrate for end-repair and ligation with barcoded Y-adapters. Prior to ligation, the A-tailing step is performed at 58 °C to preserve sub-nucleosomal fragments in the library (N. Liu et al., 2018; Neiman et al., 2012). End-repair and adapter ligation reactions were performed on a separate thermocycler. Chromatin proteins are then digested with Proteinase K, and adapter ligated DNA fragments are purified on the Biomek FX using two rounds of pre-PCR Ampure bead cleanups with size selection. PCR enrichment reactions are performed on a thermocycler using the KAPA PCR kit (KAPA Cat#KK2502). Two rounds post-PCR Ampure bead cleanups with size selection are performed on the Biomek FX to remove unwanted proteins and self-ligated adapters.

The size distributions of AutoCUT&RUN libraries were analyzed on an Agilent 4200 TapeStation, and library yield was quantified by Qubit Fluorometer (Thermo Fisher). Up to 24 barcoded AutoCUT&RUN libraries were pooled per lane at equimolar concentration for paired-end 25 × 25 bp sequencing on a 2-lane flow cell on the Illumina HiSeq 2500 platform at the Fred Hutchinson Cancer Research Center Genomics Shared Resource.

Antibodies

We used Rabbit anti-CTCF (1:100, Millipore Cat#07-729), Rabbit anti-NPAT (1:100, Termo Fisher Cat#PA5-66839), Rabbit anti-H3K4me1 (1:100, Abcam Cat#ab8895), Rabbit anti-H3K4me2 (1:100, Millipore Cat#07-030), Rabbit anti-H3K4me3 (1:100, Active Motif Cat#39159), Rabbit anti-H3K27me3 (1:100, Cell Signaling Tech Cat#9733S). Since pA-MNase does not bind efficiently to many mouse antibodies, we used a rabbit anti-Mouse IgG (1:100, Abcam, Cat#ab46540) as an adapter. H3K27ac was

profiled by AutoCUT&RUN in H1 and K562 cells and manually in VUMC-10 and SU-DIPG-XIII cell lines using Rabbit anti-H3K27ac (1:50, Millipore Cat#MABE647). H3K27ac was profiled by AutoCUT&RUN in VUMC-10 and SU-DIPG-XIII cell lines and xenografts using Rabbit anti-H3K27ac (1:100, Abcam Cat#ab45173).

Cell culture

Human K562 cells were purchased from ATCC (Manassas, VA, Catalog #CCL-243) and cultured according to supplier's protocol. H1 hESCs were obtained from WiCell (Cat#WA01-lot#WB35186) and cultured in Matrigel™ (Corning) coated plates in mTeSR™1 Basal Media (STEMCELL Technologies cat# 85851) containing mTeSR™1 Supplement (STEMCELL Technologies cat# 85852). Pediatric DMG cell lines VUMC-DIPG-10 (Esther Hulleman, VU University Medical Center, Amsterdam, Netherlands) and SU-DIPG-XIII (Michelle Monje, Stanford University, CA) were obtained with material transfer agreements from the associated institutions. Cells were maintained in NeuroCult NS-A Basal Medium with NS-A Proliferation Supplement (STEMCELL Technologies, cat# 05751), 100 U/mL of penicillin/streptomycin, 20 ng/mL epidermal growth factor (PeproTech, cat# AF-100-15), and 20 ng/mL fibroblast growth factor (PeproTech, cat# 100-18B).

Patient-derived xenografts

All mouse studies were conducted in accordance with Institute of Animal Care and Use Committee-approved protocols. NSG mice were bred in house and aged to 2–3 months prior to tumor initiation. Intracranial xenografts were established by stereotactic injection of 100,000 cells suspended in 3 µL at a position of 2 mm lateral and 1 mm posterior to lambda. Symptomatic mice were euthanized and their tumors resected for analysis and snap-frozen for storage. To prepare xenograft samples for AutoCUT&RUN, the tissue was thawed and pipetted up-and-down in CUT&RUN wash buffer to break up clumps before adding concanavalin A-coated magnetic beads.

Annotation and data analysis

We aligned paired-end reads using Bowtie2 version 2.2.5 with options: local—very-sensitive-local—no-unal—no-mixed—no-discordant—phred33 -I 10 -X 700. For mapping spike-in fragments, we also used the—no-overlap—no-dovetail options to avoid cross-mapping of the experimental genome to that of the spike-in DNA (Langmead & Salzberg, 2012). Files were processed using bedtools and UCSC bedGraphToBigWig programs (Kent et al., 2010; Quinlan & Hall, 2010).

To examine correlations between the genome-wide distributions of various samples, we generated bins of 500 bp spanning the genome, creating an array with approximately 6 million entries. Reads in each bin were counted, and the log2-transformed values of these bin counts were used to determine a Pearson correlation score between different experiments. Hierarchal clustering was then performed on a matrix of the Pearson scores.

To examine the distribution of histone mark profiles around promoters, a reference list of genes for build hg19 were downloaded from the UCSC table browser (https://genome.ucsc.edu/cgi-bin/hgTables) and oriented according to the directionality of gene transcription for further analysis. Genes with TSSs within 1 kb of each other were removed, as were genes mapping to the mitochondrial genome, creating a list of 32,042 TSSs. RNA-sequencing data were obtained from the ENCODE project for H1 and K562 cells (ENCSR537BCG and ENCSR000AEL). RNA reads were counted using featureCounts (http://bioinf.wehi.edu.au/featureCounts/) and converted to Fragments Per Kilobase per Million mapped reads (FPKM) and assigned to the corresponding TSS as a gene expression value. ATAC-sequencing data for H1 cells were obtained from Gene Expression Omnibus (GEO) (GSE85330) and mapped to hg19 using Bowtie2. Mitochondrial DNA accounted for ~ 50% of the ATAC-seq reads and was removed in this study.

All heat maps were generated using DeepTools (Ramirez et al., 2016). All of the data were analyzed using either bash or python. The following packages were used in python: Matplotlib, NumPy, Pandas, Scipy, and Seaborn.

Training the linear regression model

To ensure the accuracy of fitting histone modification data at promoters to RNA-seq values, genes with more than one promoter were removed from the previously generated TSS list. The genes RPPH1 and RMRP were expressed at extremely high levels in H1 cells and so were considered to be outliers and were removed to avoid skewing the regression, leaving a list of n = 12,805 genes.

To assign a relative CUT&RUN signal to promoters for each histone mark, denoted by C, base pair read counts ± 1 kb of the TSS were normalized by both sequencing depth over the promoters being scored and the total number of promoters examined. The prior normalization is to account for both sequencing depth and sensitivity differences among antibodies, and the latter normalization is included so that the model can be applied to different numbers of cis-regulatory elements without changing the relative weight of each element. FPKM values were used for RNA-seq.

The linear regression model was trained to fit profiles of histone marks to RNA expression values as previously described (Karlic et al., 2010).Briefly, we used a linear combination of histone data fitted to the RNA-seq expression values: y=C1x1+⋯+$Cnxn$, where Ci is the weight for each histone modification and xi is denoted by xi=ln(Ci+αi), where C is the normalized base pair counts described above and α is a pseudo-count to accommodate genes with no expression. The RNA-seq values were similarly transformed as yi=ln($FPKMi$+αy,i). Logarithmic transformations were used to linearize the data. A minimization step was then performed to calculate pseudo-counts and weights for each histone modification that would maximize a regression line between CUT&RUN data and RNA-seq.

We expected that the histone marks H3K27ac, H3K27me3, and H3K4me2 would provide the least redundant information. The optimized three histone mark model for K562 cells is described by:

$\ln(y+0.0078)=0.858\ln(C\text{H3K27ac}+0.058)-0.615\ln(C\text{H3K27me3}+0.0816)+1.609\ln(C\text{H3K4me2}+0.054)$. This equation was used to generate all chromatin activity scores.

Calling chromatin domain for overlap analysis

To compare the global chromatin landscape of H1 and K562 cells chromatin domains were called using a custom script that enriched for regions relative to an IgG CUT&RUN control. Enriched regions among marks were compared and overlaps were identified by using bedtools intersect. Overlapping regions were quantified by the number of common base pairs and these were used to generate the Venn diagrams.

Venn diagrams

All Venn diagrams were generated using the BaRC webtool, publicly available from the Whitehead Institute (http://barc.wi.mit.edu/tools/venn/).

Calculating cell-type-specific promoter activity scores

Raw promoter chromatin scores generally fall within a range from −10 to 10, where a smaller number is indicative of less transcriptional activity. To account for outliers in the data when comparing different cell types, promoter scores within 2 standard deviations were z-normalized. Negative and zero values complicate calculating fold change, so the data were shifted in the x and y directions by the most negative values. The fold difference between promoter scores for various cell types was calculated by dividing the inverse log10-normalized promoter scores against each other. A conservative twofold cutoff was used to determine cell-type-specific promoters in each case (Figures 2.4b–e and Figures 2.7a, b). Each list of genes was classified by gene ontology (http://geneontology.org/) to identify statistically enriched biological processes.

To examine the relative similarities between cell types based on their promoter activities, scores for all promoters ≥1 kb apart were used to generate an array, and Spearman correlations were calculated for each pair-wise combination of the samples. Hierarchal clustering of the Spearman correlation values was used to visualize the relative similarities between cell types.

Peak calling on AutoCUT&RUN and ATAC-seq data

Biological replicates profiled by AutoCUT&RUN were highly correlated (Figure 2.1b), so replicates were joined prior to calling peaks. The tool MACS2 was used to call peaks (Y. Zhang et al., 2008). Replicates were joined prior to calling peaks. The tool MACS2 was used to call peaks, and the following command was used on the command line: “macs2 callpeak -t file -f BEDPE -n name -q 0.01 –keep-dup all -g 3.137e9.” An FDR cutoff of 0.01 was used.

Calculating cell-type-specific enhancer activity scores

To assemble a list of distal cis-regulatory elements in the human genome, we used MACS2 to call peaks on H3K4me2 profiles from each of our samples using the same flags described in the “Peak calling on AutoCUT&RUN and ATAC-seq” methods section. To distinguish between TSSs and putative enhancers, peaks <2.5 kb a≥way from an annotated TSS were removed, and windows ±1 kb around these putative enhancers were assigned chromatin activity scores using the algorithm trained to predict promoter activity. Correlation matrices comparing the enhancer scores between samples were generated in the same manner as the correlation matrix comparing promoter scores between samples.

Data access

All data generated and used in this manuscript have been deposited in GEO: GSE120011.

Author Contribution

DHJ and SH optimized the AutoCUT&RUN protocol. DHJ, JFS, KA, and SH designed experiments. DHJ performed experiments with the help of CHM and JMO who obtained the DMG cell lines and prepared the patient-derived xenograft samples. DHJ, SJW, and MPM developed algorithms and analyzed the data. DHJ, KA, and SH wrote the manuscript. All authors read and approved the final manuscript.

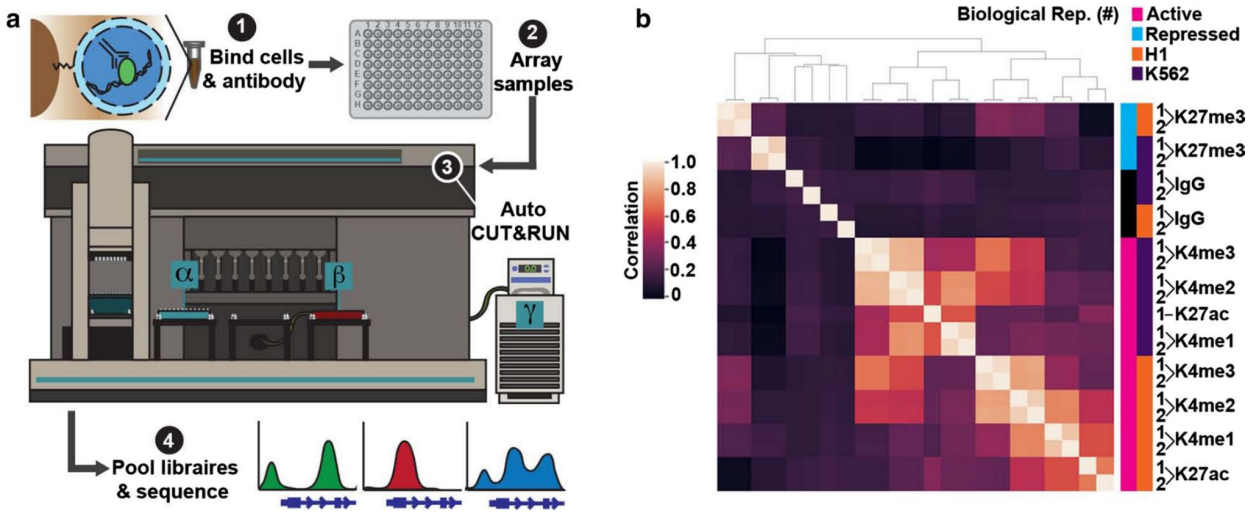


Figure 2.1: An automated platform for high-throughput in situ profiling of chromatin proteins.

a) AutoCUT&RUN workfow. (1) Cells or tissue are bound to concanavalin A-coated beads, permeabilized with digitonin, and incubated with an antibody targeting a chromatin protein. (2) Samples are arrayed in a 96-well plate and (3) processed on a Biomek robot ftted with a 96-well magnetic plate for magnetic separation during washes (α), and an aluminum chiller block (β) routed to a circulating water bath (γ) for temperature control. (4) AutoCUT&RUN produces in 2 days up to 96 libraries that are ready to be pooled and sequenced. b) Hierarchically clustered correlation matrix of AutoCUT&RUN profles of histone-H3 modifcations that mark active (pink) and repressed (blue) chromatin in H1 (orange) and K562 (purple) cells. Pearson correlations were calculated using the log2-transformed values of read counts split into 500 bp bins across the genome.

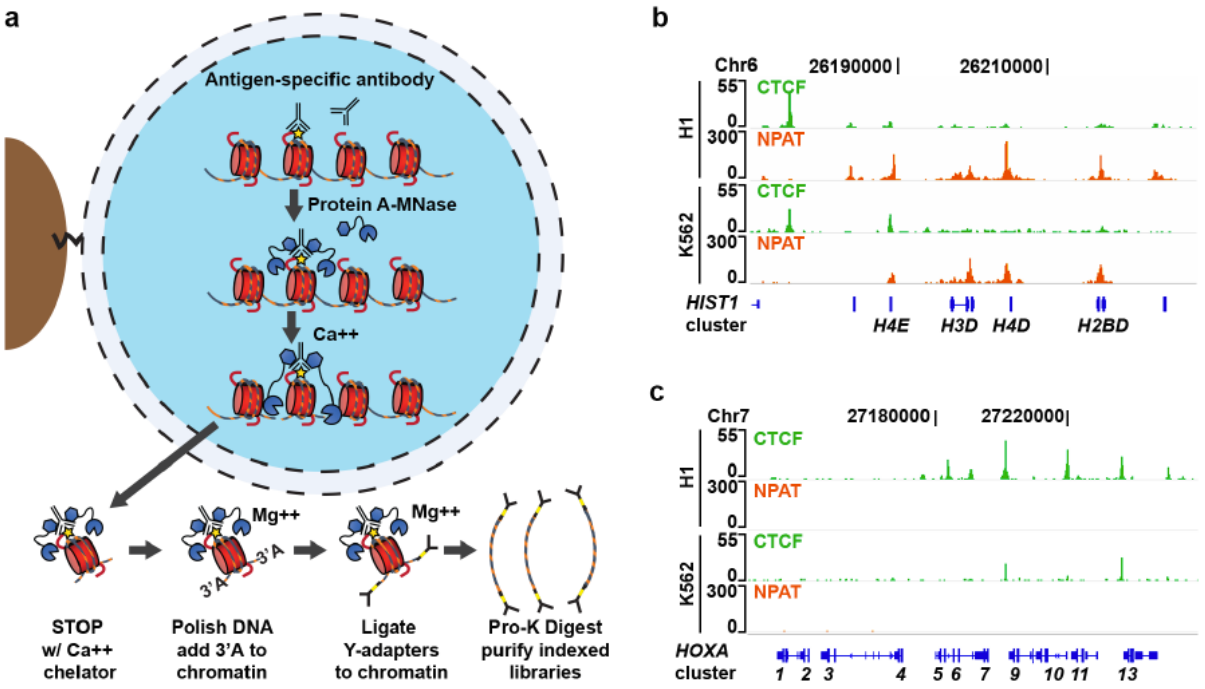


Figure 2.2: AutoCUT&RUN accurately maps NPAT and CTCF.
a) A modified CUT&RUN protocol allows for automation. ConA bead-bound samples are incubated with a chromatin protein-specific antibody, and arrayed on the Biomek for successive washes, tethering of a proteinA-MNase fusion protein, and cleavage of DNA by adding Ca2+. To avoid purifying the digested DNA prior to library prep, the reaction is stopped with an EGTA only STOP buffer which specifically chelates Ca2+ while leaving adequate Mg2+ to allow End-polishing and Ligation of Illumina Y-adapters to the chromatin fragments. Chromatin protein is then digest with Proteinase-K and the indexed CUT&RUN libraries are purified on the Biomek using Ampure Magnetic Beads. b) Genome browser tracks of NPAT and CTCF AutoCUT&RUN showing NPAT enrichment at promoters of the HIST1 gene cluster in both H1 and K562 cells. c) Genome browser tracks confirming CTCF is bound to insulator regions in the HOXA locus.

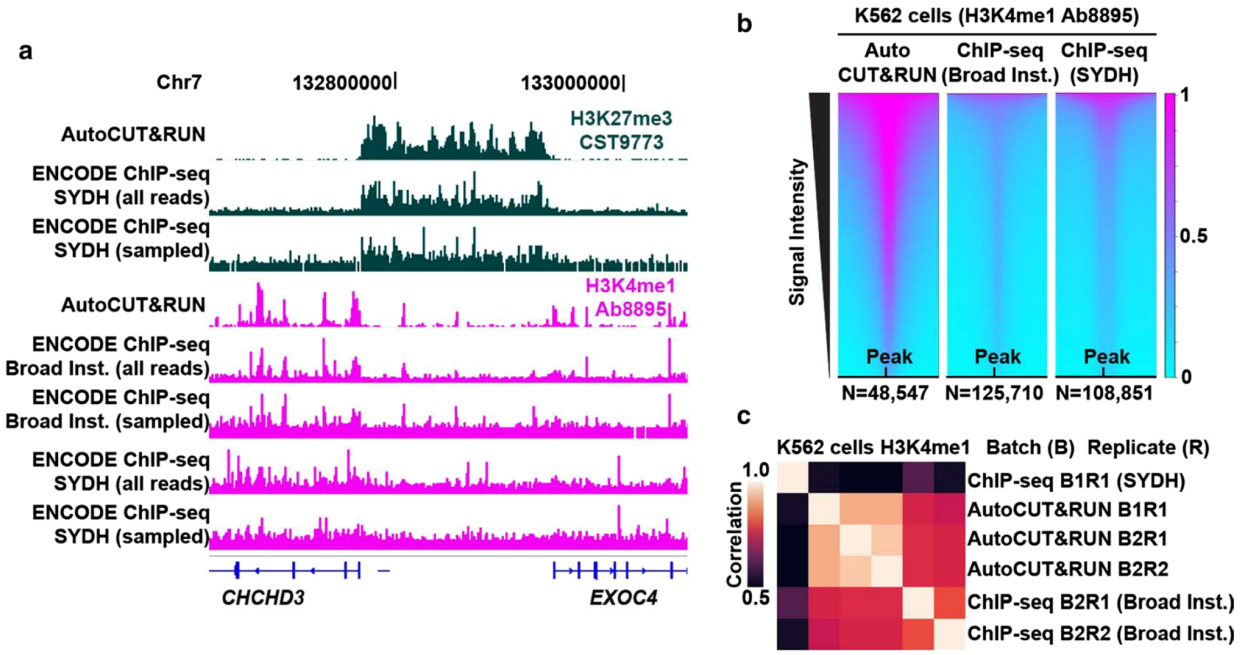


Figure 2.3: Comparison of AutoCUT&RUN versus ENCODE ChIP-seq.
A representative silenced domain fanked by active genes is shown for human K562 cells probed using the same antibodies by either AutoCUT&RUN or ChIP-seq. For each comparison, ChIP-seq tracks are shown that include either all sequenced fragments or the same number of sampled fragments as the AutoCUT&RUN data. b Heat map comparison of H3K4me1 Ab8895 ±1 kb around peak centers (summits) called by MACS2 for AutoCUT&RUN data and for ENCODE ChIP-seq data (broad peaks by Broad Institute and narrow peaks by SYDH). Signal intensities refect the relative number of reads that fall within peaks. c Correlation matrix comparing batches (B) and replicates (R) between AutoCUT&RUN and ENCODE ChIP-seq datasets for H3K4me1 Ab889

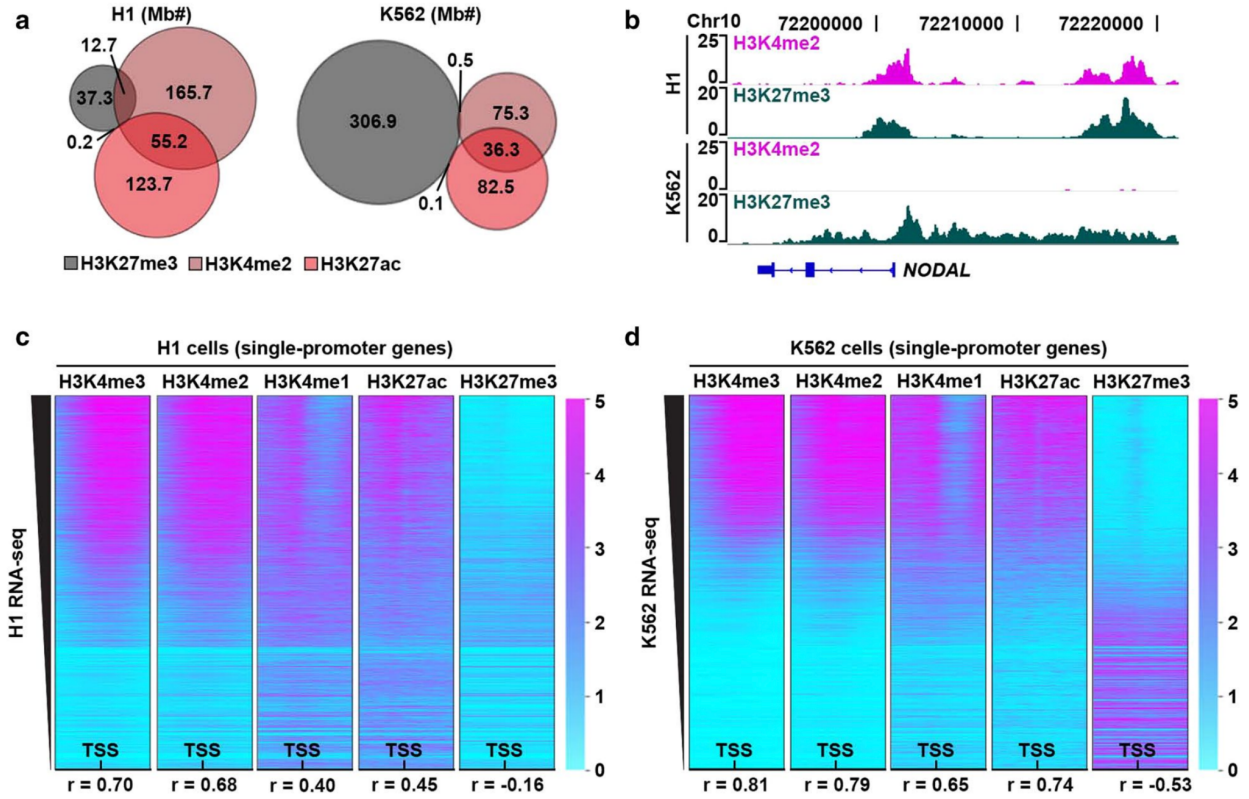


Figure 2.4: AutoCUT&RUN reproduces the expected chromatin landscape of H1 and K562 cells. a) Scaled Venn diagrams showing the relative amount of the genome that falls within H3K27me3 (gray), H3K4me2 (brown), and H3K27ac (red) domains in H1 cells and K562 cells. Numbers indicate megabases (Mb). b) Genome browser tracks showing the overlap of H3K4me2 and H3K27me3 in H1 cells, as well as the expansion of H3K27me3 domains and loss of overlap with H3K4me2 in K562 cells at a representative locus (NODAL). c) Heat maps showing the distribution of AutoCUT&RUN profles of histone modifcations in H1 cells centered on the TSSs of genes with a single promoter, oriented left-to-right according to the 5′-to-3′ direction of transcription and rank-ordering according to RNA-seq values (FPKM). d) Heat maps showing the distribution of AutoCUT&RUN histone modifcation profles on transcriptionally active and repressed promoters in K562 cells. Pearson correlations (r value) between AutoCUT&RUN profles of individual histone marks around these TSSs and their corresponding RNA-seq values are indicated

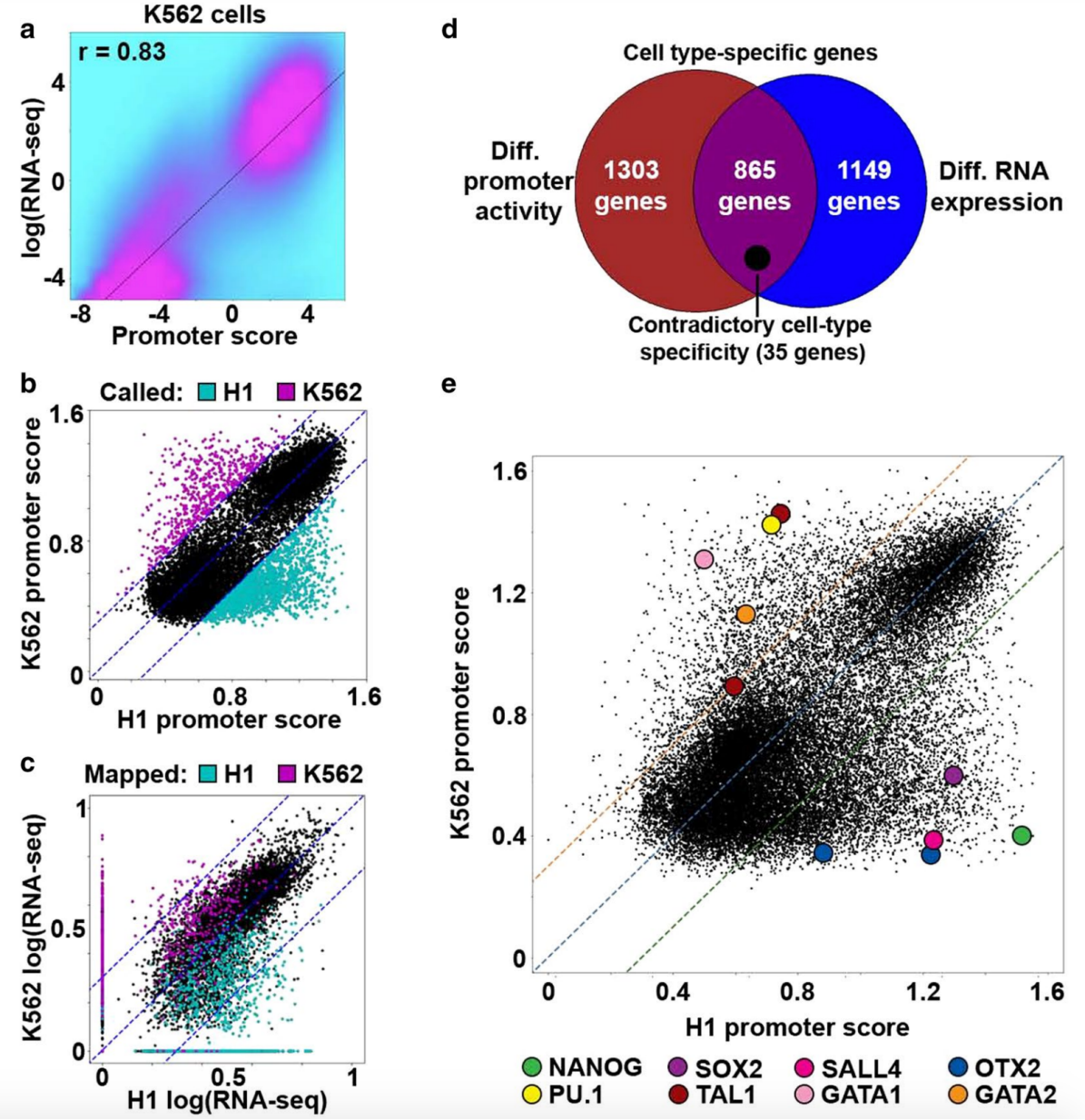


Figure 2.5: A linear regression model accurately predicts cell-type specifc promoter activity.
a) Density scatterplot comparing RNA-seq values for single-promoter genes to K562 promoter scores predicted by the model trained on K562 data. b) Scatterplot of promoter chromatin scores for single-promoter genes in H1 and K562 cells. Colored dots indicate that the promoter scores are ≥twofold enriched in either H1 cells (cyan) or K562 cells (magenta). c) Scatterplot of promoter scores that are ≥twofold enriched in either H1 cells (cyan) or K562 cells (magenta) mapped onto their corresponding RNA-seq values. Blue dotted lines indicate the twofold diference cutoff. d) Scaled Venn diagram showing the overlap between genes called as cell type specifc according to their promoter scores, or according to their RNA expression values. Genes predicted to have contradictory cell type specifcities according to

promoter activity modeling versus RNA-seq are indicated (scaled black circle). e) Scatterplot comparing the H1 and K562 scores of all promoters separated by ≥2 kb. Master regulators of H1 and K562 cell identities are indicated as colored circles. Both OTX2 and TAL1 have two promoters that can be distinguished

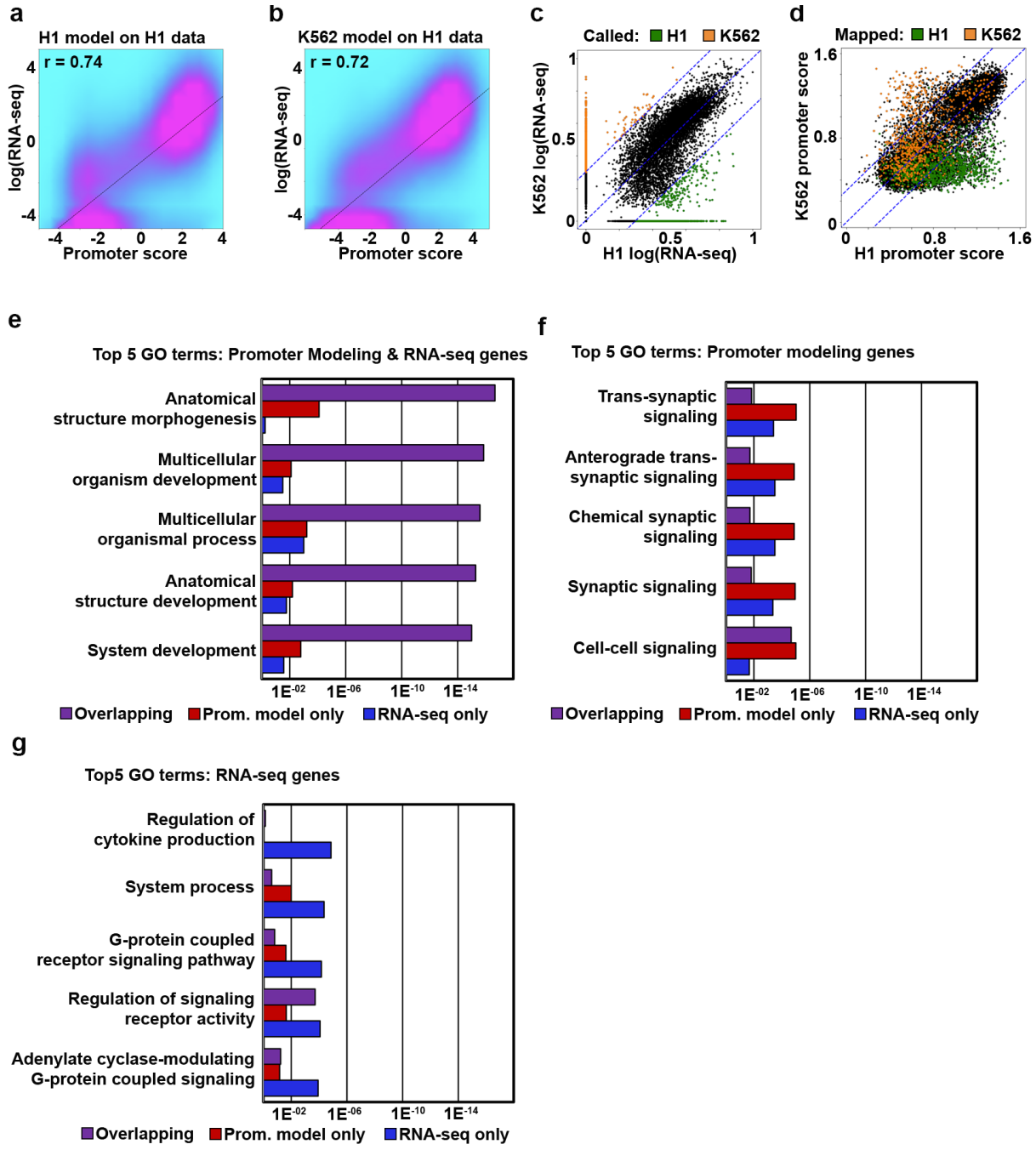


Figure 2.6: Developing a linear regression model to predict the activity of cis-regulatory elements.

a) Density scatterplot comparing H1 RNA-seq values for single promoter genes to H1 promoter scores predicted by the model trained on H1 data. b) Density scatterplot comparing H1 RNA-seq values for single-promoter genes to H1 promoter scores predicted by the model trained on K562 data. c) Scatterplot of RNA-seq values for single-promoter genes in H1 and K562 cells. Colored dots indicate the RNA expression levels are >2-fold enriched in either H1 cells (green) or K562 cells (orange). d) Scatterplot showing the distribution of genes with RNA-seq values that are >2-fold enriched in either H1 cells (green) or K562 cells (orange) mapped onto their corresponding promoter chromatin scores. Blue dotted lines indicated the 2-fold difference cut-off. e) Bar graph showing the top five Gene Ontology (GO) terms overrepresented in the collection of cell-type specific genes identified by both promoter activity modeling as well as RNA-seq (purple), and the relative enrichment of these terms in the collections of genes uniquely identified as cell-type specific by promoter activity modeling (red)
or RNA-seq (blue). f) Bar graph showing the top five GO terms overrepresented in the collection of genes identified as cell-type specific according to promoter modeling scores only. g) Bar graph showing the top five GO terms overrepresented in the collection of genes uniquely identified as cell-type specific according to RNA-seq.

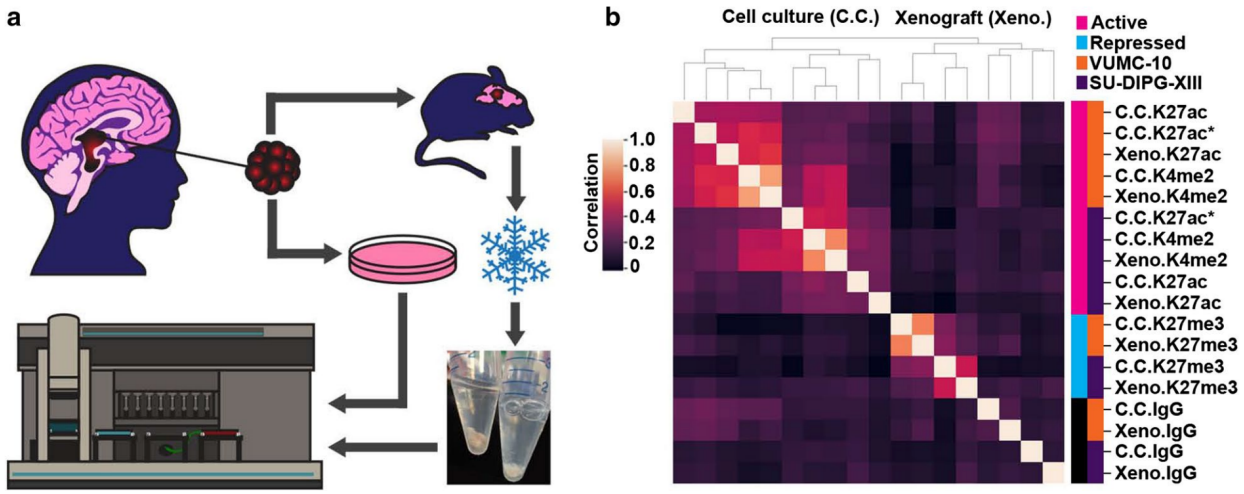


Figure 2.7: AutoCUT&RUN is suitable for profling the chromatin landscape of frozen tumor samples. a) DMG experimental setup. Two DMG cell lines derived from a similar region of the brainstem were grown as xenografts in the brains of immunocompromised mice, and upon forming tumors were resected and frozen. Xenografts were thawed and processed by AutoCUT&RUN in parallel with control DMG samples harvested directly from cell culture. b) Hierarchically clustered correlation matrix of AutoCUT&RUN profles of histone-H3 modifcations that mark active (pink) and repressed (blue) chromatin in VUMC-10 (orange) and SU-DIPG-XIII (purple) cells grown in cell culture (C.C.) or as xenografts (Xeno.). As a quality control, H3K27ac was also profled manually in these cell lines using a diferent antibody (*). Pearson correlations were calculated using the log2-transformed values of read counts split into 500 bp bins across the genome.

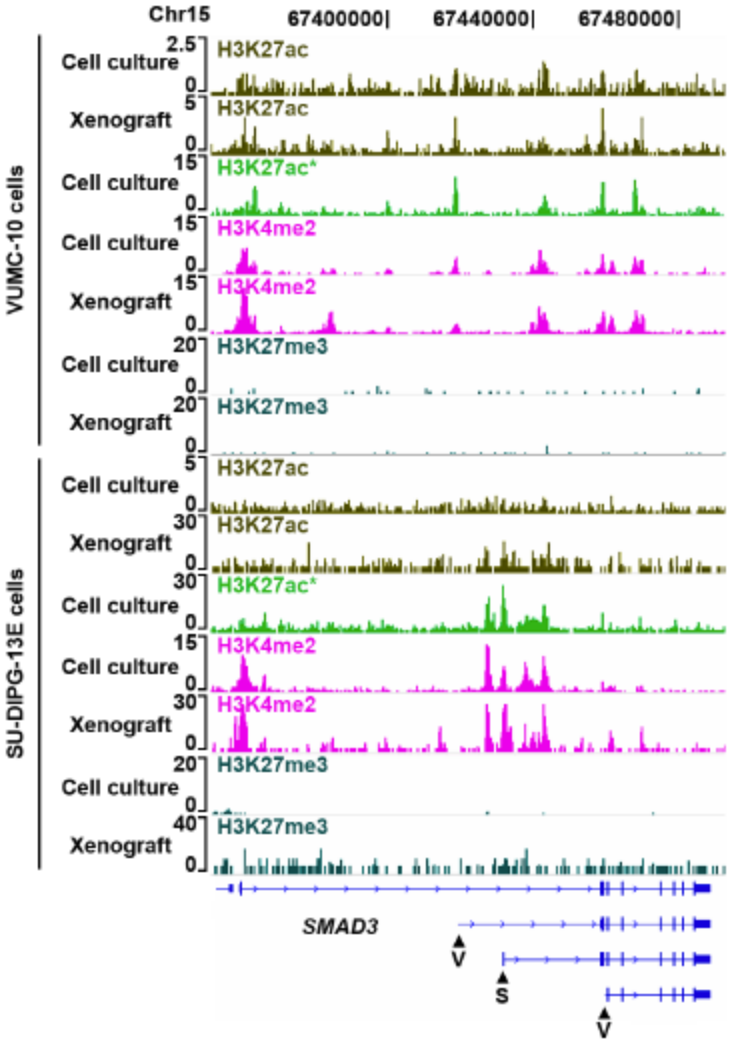


Figure 2.8: DMG subtype-specific SMAD3 promoter activities.
Genome browser tracks of histone marks profiled by AutoCUT&RUN in VUMC-10 and SU-DIPG-XIII cells at a representative locus (SMAD3) showing the concordance of profiles from cell culture and xenograft samples. The H3K27ac signal in SU-DIPG-XIII cells was
noisy, but this issue is antibody specific. For comparison H3K27ac was also profiled manually using an alternative antibody (*). Arrowheads indicate promoters that are predicted to be specifically active in VUMC-10 (V) or SUDIPG-XIII (S) cells.

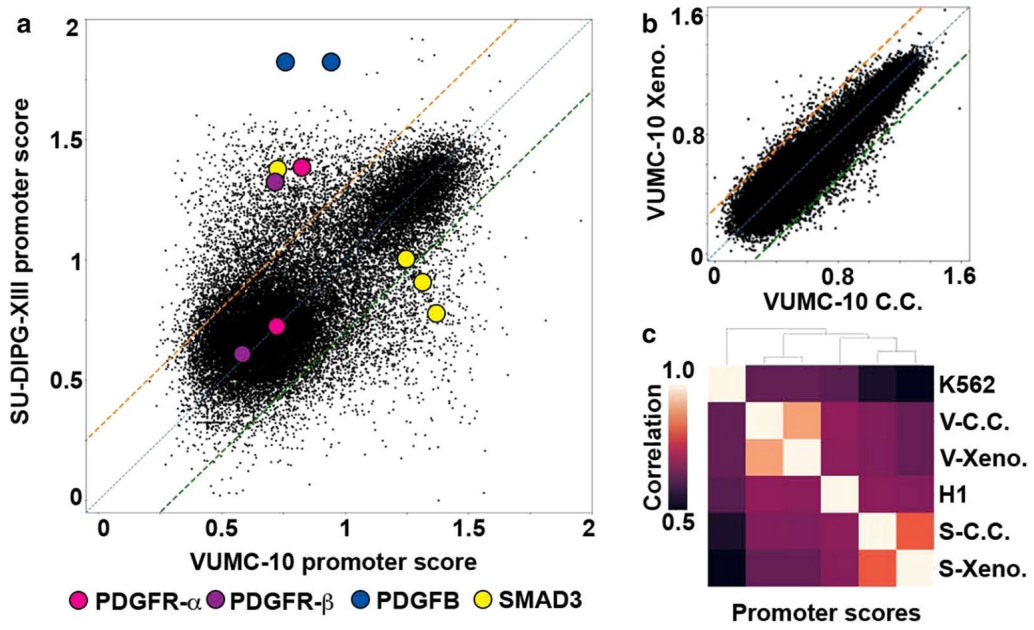


Figure 2.9: Promoter activity modeling distinguishes gene activities in DMG samples.
a) Scatterplot comparing the promoter scores of VUMC-10 and SU-DIPG-XIII cell culture samples. Locations of the promoters of several cell signaling components implicated in tumor growth are indicated as colored circles. b) Scatterplot comparing the promoter scores of VUMC-10 cell culture (C.C.) and xenograft (Xeno.) samples. Only 388 promoters have a≥twofold diference in activity modeling scores between these samples. c) Hierarchically clustered matrix of Spearman correlations of promoter chromatin scores between VUMC-10 (V) and SU-DIPG-XIII (S) cells grown in cell culture (C.C.) or as xenografts (Xeno.), as well as H1 and K562 cells.

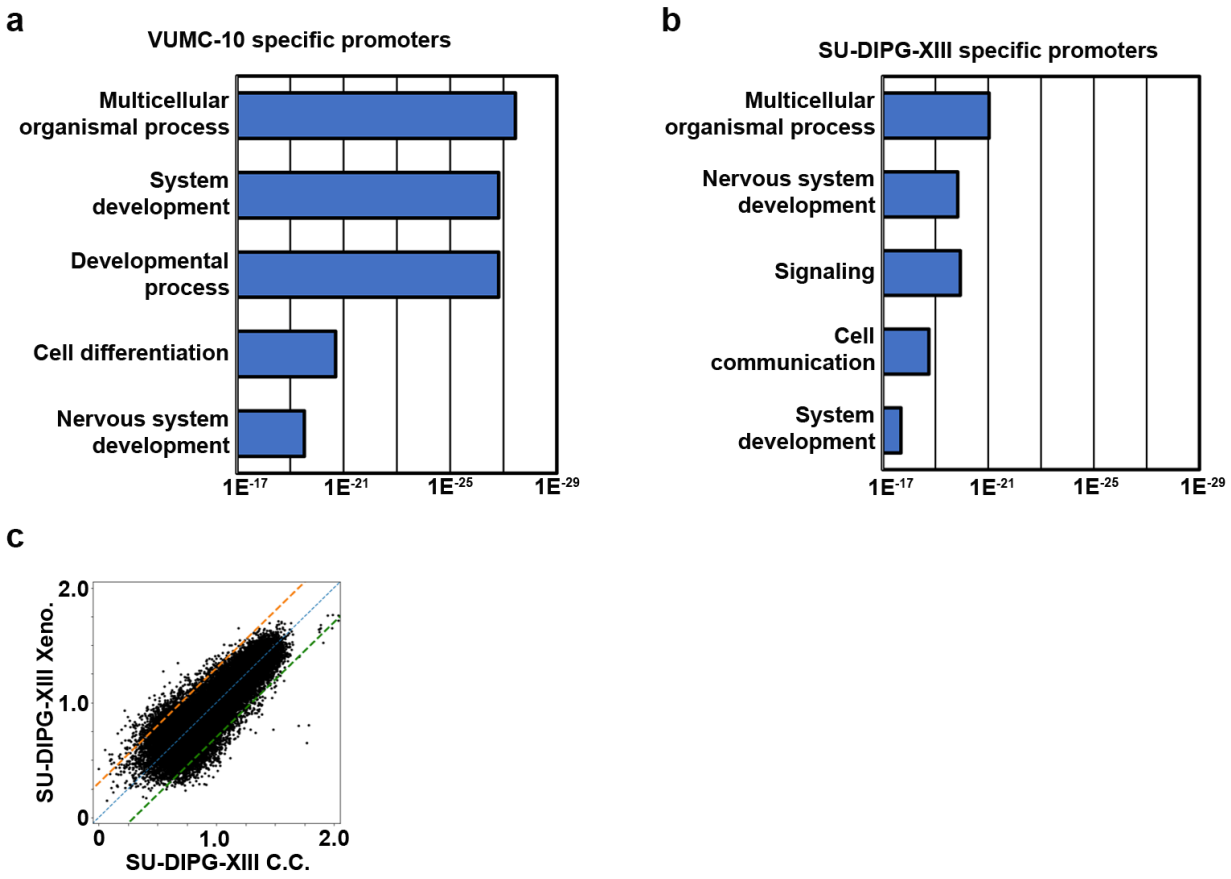
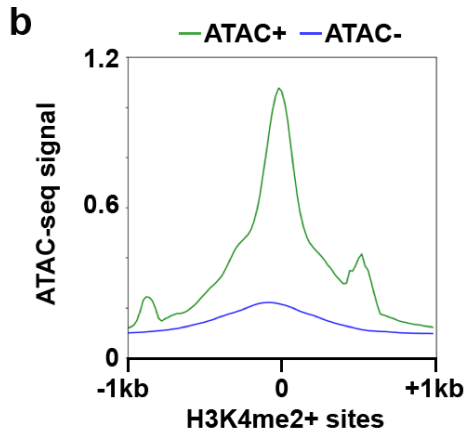


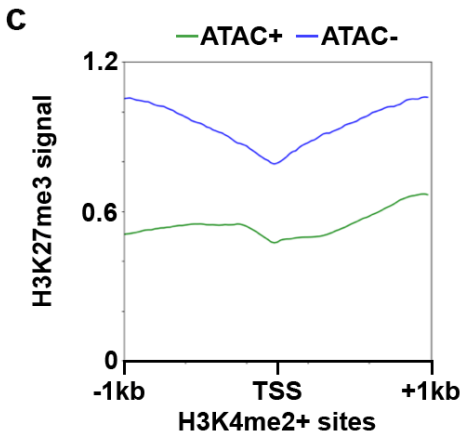
Figure 2.10: AutoCUT&RUN identifies DMG specific gene regulatory programs.
a) GO terms that are overrepresented in the collection of promoters that are >2-fold enriched in VUMC-10 cells according to promoter chromatin scores. b) GO terms that are overrepresented in the collection of promoters that are >2-fold enriched in SUDIPG-XIII cells according to promoter chromatin scores. c) Scatterplot comparing the promoter scores of SU-DIPG-XIII cell culture (C.C.) and xenograft (Xeno.) samples. 1,619 promoters have a >2-fold difference in promoter chromatin scores between these samples.

a

Overlap with ATAC in H1 cells

mark	# peaks	# overlap
ATAC	52,270	N/A
H3K27ac	38,053	8,992
H3K27me3	21,160	4,813
H3K4me1	146,010	23,242
H3K4me2	108,122	36,725
H3K4me3	47,294	31,672





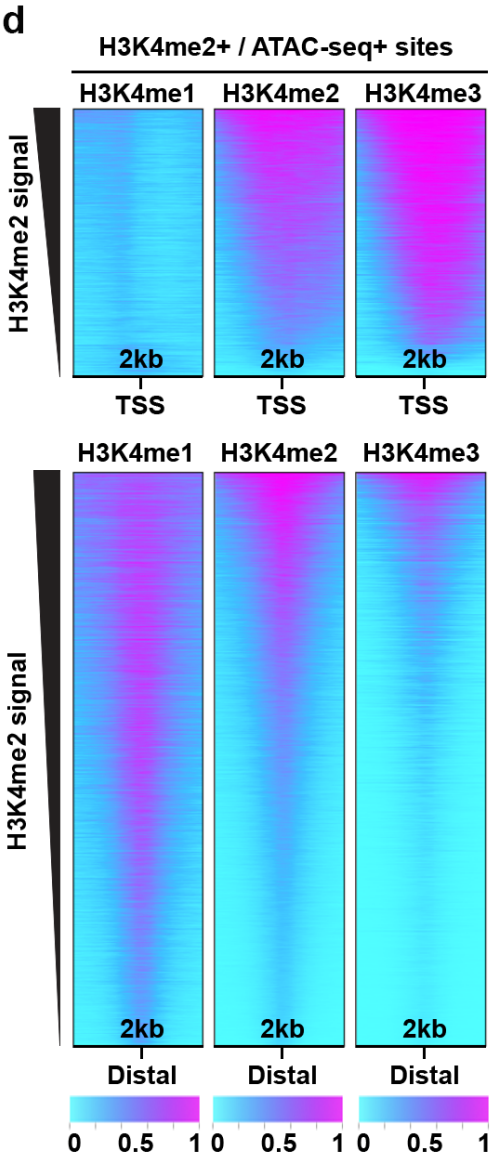


Figure 2.11: AutoCUT&RUN is a sensitive method to distinguish proximal and distal cisregulatory elements.

a) Table of the overlap of accessible chromatin sites (ATACseq peaks) and peaks called on various AutoCUT&RUN profiles of histone marks in H1 cells. b) Mean enrichment of ATAC signal at H3K4me2 peaks that were either called as ATAC+ (green) or ATAC- (blue). c) Mean enrichment of H3K27me3 signal at H3K4me2+ TSSs that were either called as ATAC+ (green) or ATAC- (blue). d Heat maps showing the distribution of normalized H3K4me1, H3K4me2 and H3K4me3 profiles over all H3K4me2+/ATAC+ TSSs and distal regulatory elements (Distal).

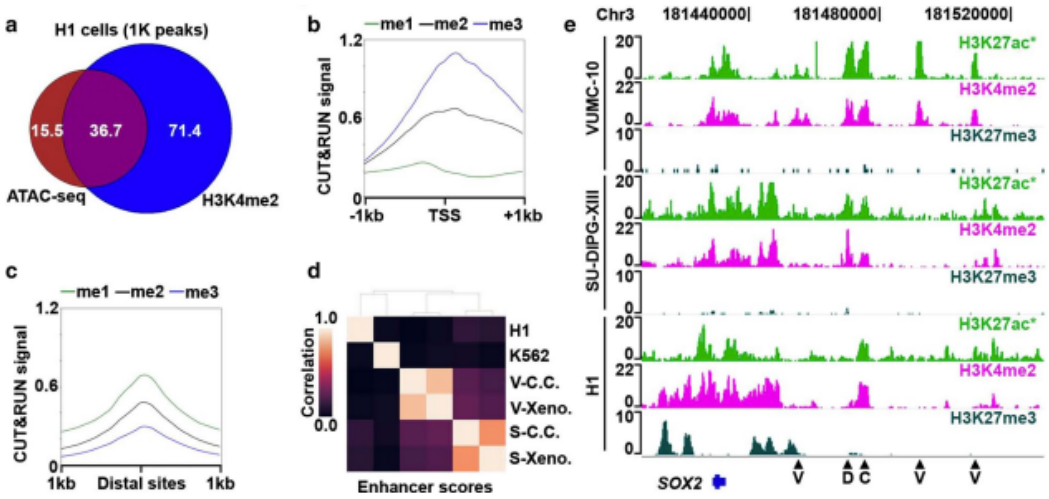


Figure 2.12: AutoCUT&RUN identifies cell-type specific enhancer elements.
a) Scaled Venn diagram showing the overlap of accessible chromatin sites (ATAC-seq peaks) and peaks called on H3K4me2 AutoCUT&RUN profiles in H1 cells. Numbers are provided as 1 thousand peak units. b) Mean enrichment of H3K4me1 (green) H3K4me2 (black) and H3K4me3 (blue) over all H3K4me2+/ATAC+TSSs. c) Mean enrichment of H3K4me1 (green) H3K4me2 (black) and H3K4me3 (blue) over all H3K4me2+/ATAC+distal sites. d) Hierarchically clustered matrix of Spearman correlations of enhancer chromatin scores in VUMC-10 (V) and SU-DIPG-XIII (S) cells grown in cell culture (C.C.) or as xenografts (Xeno.), as well as H1 and K562 cells. e) Genome browser tracks showing the location of putative enhancer elements (arrow heads) that are specific to VUMC-10 cells (V), both DMG cell lines (D), or common to DMG cells and H1 cells (C) at a representative locus (SOX2). In VUMC-10 and SU-DIPG-XIII cells, the H3K27ac track that is shown was also profiled manually as a quality control (*)

Chapter 3

CUT&Tag for efficient epigenomic profiling of small samples and single cells

Modified from an article originally published in *Nature Communications:*
Hatice S. Kaya-Okur, Steven J. Wu, Christine A. Codomo, Erica S. Pledger, Terri D. Bryson, Jorja G. Henikoff, Kami Ahmad & Steven Henikoff. *Nature Communications* volume 10, Article number: 1930 (2019)

Abstract

Many chromatin features play critical roles in regulating gene expression. A complete understanding of gene regulation will require the mapping of specific chromatin features in small samples of cells at high resolution. Here we describe Cleavage Under Targets and Tagmentation (CUT&Tag), an enzyme-tethering strategy that provides efficient high-resolution sequencing libraries for profiling diverse chromatin components. In CUT&Tag, a chromatin protein is bound in situ by a specific antibody, which then tethers a protein A-Tn5 transposase fusion protein. Activation of the transposase efficiently generates fragment libraries with high resolution and exceptionally low background. All steps from live cells to sequencing-ready libraries can be performed in a single tube on the benchtop or a microwell in a high-throughput pipeline, and the entire procedure can be performed in one day. We demonstrate the utility of CUT&Tag by profiling histone modifications, RNA Polymerase II and transcription factors on low cell numbers and single cells.

Introduction

The advent of massively parallel sequencing and the dramatic reduction in cost per base has fueled a genomics revolution, however, the full promise of epigenomic profiling has lagged owing to limitations in methodologies used for mapping chromatin fragments to the genome (Zentner & Henikoff, 2014). Chromatin immunoprecipitation with sequencing (ChIP-seq) and its variations (Kasinathan et al., 2014; Policastro & Zentner, 2018; Rhee & Pugh, 2011; Skene & Henikoff, 2017) suffer from low signals, high backgrounds and epitope masking due to cross-linking, and low yields require large numbers of cells (Policastro & Zentner, 2018; Teytelman et al., 2013). Alternatives to ChIP include enzyme-tethering methods for unfixed cells, such as DamID7, ChEC-seq (Zentner et al., 2015), and CUT&RUN (Skene et al., 2018; Skene & Henikoff, 2017) where a specific protein of interest is targeted in situ and then profiled genome-wide. For example, CUT&RUN, which is based on Laemmli's Chromatin ImmunoCleavage (ChIC) strategy (Schmid et al., 2004a), maps a chromatin protein by successive binding of a specific antibody, and then tethering a Protein A/Micrococcal Nuclease (pA-MNase) fusion protein in permeabilized cells without cross-linking (Skene & Henikoff, 2017). MNase is activated by addition of calcium, and fragments are released into the supernatant for extraction of DNA, library preparation and

paired-end sequencing. CUT&RUN provides base-pair resolution of specific chromatin components with background levels that are much lower than with ChIP-seq, dramatically reducing the cost of genome-wide profiling. Although CUT&RUN can generate high-quality data from as few as 100–1000 cells, it must be followed by DNA end polishing and adapter ligation to prepare sequencing libraries, which increases the time, cost and effort of the overall procedure. Moreover, the release of MNase-cleaved fragments into the supernatant with CUT&RUN is not well-suited for application to single-cell platforms (*Mezger, A. et al. High-Throughput Chromatin Accessibility Profiling at Single-Cell Resolution. Nat. Commun.9, 3647 (2018).*, n.d.; Zheng, 2017).

Here we overcome the limitations of ChIP-seq and CUT&RUN using a transposome that consists of a hyperactive Tn5 transposase (Picelli, 2014; Reznikoff, 2003)—Protein A (pA-Tn5) fusion protein loaded with sequencing adapters. Tethering in situ followed by activation of pA-Tn5 results in factor-targeted tagmentation, generating fragments ready for PCR enrichment and DNA sequencing. Beginning with live cells, Cleavage Under Targets and Tagmentation (CUT&Tag) provides amplified sequence-ready libraries in a day on the bench top or in a high-throughput format. We show that a variety of chromatin components can be profiled with exceptionally low backgrounds using low cell numbers and even single cells. This easy, low-cost method will empower epigenetic studies in diverse areas of biological research.

Results

Efficient profiling of nucleosomes and RNAPII with CUT&Tag

To implement chromatin profiling by tagmentation (Figure 3.1a), we incubated intact permeabilized human K562 cells with an antibody to lysine-27-trimethylation of the histone H3 tail (H3K27me3), an abundant histone modification that marks silenced chromatin regions. We incubated cells with a secondary anti-rabbit antibody to increase the local concentration of antibody bound on chromatin sites, then incubated cells with an excess of pA-Tn5 fusion protein pre-loaded with sequencing adapters to tether the enzyme at antibody-bound sites in the nucleus. The transposome has inherent affinity for exposed DNA (Buenrostro et al., 2013; Steiniger et al., 2006), and so we washed cells under stringent conditions to remove un-tethered pA-Tn5. We then activated the transposome by addition of Mg++, integrating adapters spanning sites of H3K27me3-containing nucleosomes. Finally, fragment libraries were enriched from purified DNA and pooled for multiplex paired-end sequencing on an Illumina HiSeq flow-cell. The entire protocol manipulates all steps in a single reaction tube (Figure 3.1b), where permeabilized cells are first mixed with an antibody, and then immobilized on Concanavalin A-coated paramagnetic beads, allowing magnetic handling of the cells in all successive wash and reagent incubation steps. For standardization between experiments, we used the small amount of tracer genomic DNA derived from the E. coli during transposase protein production to normalize sample read counts in lieu of the heterologous spike-in DNA that is recommended for CUT&RUN9 (Figure 3.2a).

Display of ~8 million reads mapped to the human genome assembly shows a clear pattern of large chromatin domains marked by H3K27me3 (Figure 3.3a). We also obtained profiles for H3K4me1 and H3K4me2 histone modifications, which mark active chromatin sites. In contrast, incubation of cells with a non-specific IgG antibody, which measures untethered integration of adapters, produced very sparse landscapes (Figure 3.3a). To assess the signal-to-noise of CUT&Tag relative to other methods we compared it with profiling generated by CUT&RUN (Janssens, 2018) and by ChIP-seq (Landt et al., 2012) for the same H3K27me3 rabbit monoclonal antibody in K562 cells. To directly compare the three techniques, we set the read depth of each dataset to 8 million reads each. Landscapes for each of the three methods are similar, but background noise dominates in ChIP-seq datasets (Figure 3.3a), and it thus appears that ChIP-seq will require substantially greater read depth to distinguish chromatin features from background. In contrast, both CUT&RUN and CUT&Tag profiles have extremely low background noise levels. As expected, very different profiles were seen in the same region for a different human cell type, H1 embryonic stem (H1 ES) cells (Figure 3.3b). To more quantitatively compare signal and noise levels in each method, we generated heatmaps around genomic sites called from H3K4me1 modification profiling for each method, where the same antibody had been used. After sampling each dataset to 8 million reads for comparison, we found that CUT&Tag for this histone modification shows moderately higher signals compared to CUT&RUN throughout the list of sites (Figure 3.3c). Both methods have low backgrounds around the sites. In contrast, ChIP-seq signal has a very narrow dynamic range that is ~1/20 of the CUT&Tag signal range, and much weaker signals across the majority of sites. To quantitatively compare methods, we displayed the average read counts for CUT&Tag, CUT&RUN and ChIP-seq datasets for the H3K4me1 histone mark around the top 10,000 peaks defined by MACS2 on an H3K4me1 ChIP-seq dataset (Figure 3.3g). We found that CUT&Tag profiling gives substantially more signal accumulation at these sites, implying that CUT&Tag will be most effective at distinguishing chromatin features with fewest reads.

The transcriptional status of genes and regulatory elements can be inferred from histone modification patterns, but gene expression is directly read out by profiling chromatin-bound RNA polymerase II (RNAPII). We used an antibody to the S2/S5-phosphorylation (S2/5p) forms of RNAPII, which distinguish engaged polymerase (Zaborowska et al., 2016). Landscapes show enrichment of RNAPII CUT&Tag reads at many genes (Figure 3.3a and Figure 3.4a), and a promoter heatmap reveals that this enrichment is predominantly at the 5′ ends of active genes (“An Integrated Encyclopedia of DNA Elements in the Human Genome,” 2012) (Figure 3.3d). These results were confirmed by the observation of similar CUT&Tag patterns using antibodies to S2p, S5p and S7p forms of RNAPII (Figure 3.4a and Figure 3.5a, b).

To validate RNAPII CUT&Tag without relying upon annotations, which are typically based on mapping of processed transcripts, we chose transcriptional run-on data obtained with the base-pair-resolution PRO-seq technique, which provides direct mapping of RNAPII using a method that is unrelated to chromatin profiling (Core, 2014). PRO-seq maps the position of the 5′ end of engaged RNAPII as it is

activated in situ, and is used to identify paused RNAPII just downstream of the transcriptional start site. Peaks were called from RNAPII S2/5p CUT&Tag and ordered using MACS2, and processed datasets from PRO-seq run-on for human K562 cells (SRA GSM1480327) were aligned to the peak calls. When ordered by RNAPII CUT&Tag MACS2 score, a close correspondence between PRO-seq occupancy and RNAPII-Ser2/5p CUT&Tag occupancy is seen (Figure 3.3e). Similar heat maps were obtained using antibodies to S2p, S5p, and S7p phosphorylation of the RNAPII C-terminal domain (Figure 3.5c).

CUT&Tag sensitively maps active sites in chromatin

Replicates for profiling of H3K4me1 modification by CUT&Tag are highly similar, demonstrating the reproducibility of the method (Figure 3.6a). We obtained similar reproducibility when we compared H3K27me3 CUT&Tag replicates (Figure 3.4c). In previous experiments with CUT&RUN profiling, we found that H3K4me2 histone modification landscapes, which are associated with active promoters and enhancers, resemble ATAC-seq profiles (Janssens, 2018). We therefore performed CUT&Tag using an antibody to H3K4me2. An example of H3K4me2 CUT&Tag profiling to published ATAC-seq in K562 cells23 is shown (Figure 3.3a). We found high occupancies for H3K4me2 at strong ATAC-seq peaks (Figure 3.3f), with much higher read counts (Figure 3.3h), implying that H3K4me2 profiling captures the most prominent accessible chromatin sites in the genome with greater sensitivity.

To quantify the sensitivity of H3K4me2 CUT&Tag relative to H3K4me2 CUT&RUN (Janssens, 2018), H3K4me2 ChIP-seq (Landt et al., 2012), and ATAC-seq (X. Liu, 2017), we downsampled reads from each method, and used MACS2 with default parameters to call peaks on each dataset. We then estimated the fraction of reads falling within the called peaks. We found that both CUT&RUN and CUT&Tag populate peaks more deeply than ChIP-seq or ATAC-seq, demonstrating that they have exceptionally low signal-to-noise (Figure 3.6b). In addition, CUT&Tag more rapidly populates peaks at low sequencing depths, where ~2 million reads are equivalent to 8 million for CUT&RUN (or 20 million for ChIP-seq), demonstrating the exceptionally high efficiency of CUT&Tag. Of all the methods, only CUT&Tag reaches a fraction of 0.6 within peaks. Thus, with two histone modifications (H3K4me2 and H3K27me3), we segment the chromatin landscape into both active and silenced regions, even with relatively low sequencing depths.

CUT&Tag simultaneously maps factor binding and accessible DNA

To determine if we could use CUT&Tag for mapping transcription factor binding, we tested if pA-Tn5 tethered at transcription factors can be distinguished from accessible DNA sites in the genome. We used an antibody to the NPAT nuclear factor, a transcriptional coactivator of the replication-dependent histone genes, in CUT&Tag reactions. NPAT binds only ~80 accessible sites in the histone clusters on chromosome 1 and chromosome 624, thus we can compare true binding sites with accessible sites. In NPAT CUT&Tag profiles, ~99% of read counts accumulate at the promoters of the histone genes (Figure 3.7a). By scoring sites for correspondence to published ATAC-seq data23, we found that a smaller

number of counts are distributed across accessible sites in the K562 genome (Figure 3.7b). This probably results from some un-tethered pA-Tn5 binding to exposed DNA in situ, but it is straightforward to distinguish antibody-tethered sites from accessible sites by the vast difference in read coverage (Figure 3.7c). Indeed, calling peaks by standard algorithms on NPAT CUT&Tag data generates a list of ~9000 sites that includes histone gene promoters and 10% of ATAC-defined accessible sites (Figure 3.8). While this is only a fraction of the ~54,000 accessible sites defined in K562 cells, adjusting the threshold and stringency of NPAT peak calling may improve detection.

To test if CUT&Tag is tractable for profiling more abundant transcription factor binding sites, we profiled the CCCTC-binding factor (CTCF) DNA-binding protein. For these experiments, we varied the stringency of wash buffers to assess displacement of transcription factors from chromatin. Under low-salt and medium-salt concentration conditions we observed read counts at CTCF sites detected by CUT&RUN and by ChIP-seq (Figure 3.9a), but with additional minor peaks (Figure 3.4a). These additional peaks suggest that un-tethered pA-Tn5 contributes to coverage in these experiments. To determine if true CTCF binding sites could be distinguished from accessible features by read depth, we compared the CUT&Tag read count at high-confidence CTCF sites (defined by peak-calling on CUT&RUN data (Janssens, 2018)) to the CUT&Tag read count at accessible sites (defined by peak-calling on ATAC-seq data (X. Liu, 2017)). We found that these two distributions of read counts overlap, but that of accessible sites is lower than that of CTCF sites (Figure 3.9b). Based solely on read depth, we discriminate ~5600 CTCF bound sites with a 1% false discovery rate. Comparing motif enrichment in these two classes demonstrates that the high signals correspond to CTCF motifs (E-value = 2.1×10^{-69}), and the low signals do not.

We assessed the resolution of the CUT&Tag procedure by plotting the ends of reads centered on CTCF binding sites. This shows that CUT&Tag protects a “footprint” spanning 80 bp directly over the CTCF motif (Figure 3.9c). While the segment protected from Tn5 integration is larger than the ~45 bp protected from MNase in CUT&RUN (Skene & Henikoff, 2017), this indicates that the tethered transposase produces high resolution maps of factor binding sites. Similar footprints were obtained using different salt concentration washes, although 300–500 mM salt concentrations resulted in somewhat reduced signal-to-noise (Figure 3.9c). The high resolution of CUT&Tag provides structural details of individual sites. For example, superposition of CTCF, H3K4me1, H3K4me2, H3K4me3, and ATAC mapping at a representative site reveals the relationship between accessible DNA, CTCF binding, and modified neighboring nucleosomes (Figure 3.9d).

CUT&Tag profiles low cell number samples and single cells

ChIP requires substantial cellular material, limiting its application for experimental and clinical samples. However, we and others have previously demonstrated that tethered profiling strategies like

CUT&RUN have sufficient sensitivity that profiling small cell numbers routinely becomes feasible (Hainer et al., 2018; Skene & Henikoff, 2017). Signal improvements in CUT&Tag suggest that this method may work even more efficiently with limited samples. We first tested CUT&Tag for the H3K27me3 modification across a ~1500× range of material, from 100,000 down to 60 cells. We observed very similar high-quality chromatin profiles from all experiments (Figure 3.2b), demonstrating that high data quality is still maintained with low input material. Analyzing sample and tracer DNA in these CUT&Tag series revealed that sequencing yield is proportional to the number of cells (Figure 3.2a).

CUT&Tag has the advantage that the entire reaction from antibody binding to adapter integration occurs within intact cells. The transposase and chromatin fragments remain bound together (Amini, 2014; Picelli, 2014), and thus fragmented DNA is retained within each nucleus. We developed a simple strategy to generate chromatin profiles of individual cells, which we term single-cell CUT&Tag (scCUT&Tag) (Figure 3.10a). We performed scCUT&Tag to the H3K27me3 modification on a bulk population of K562 cells, but with gentle centrifugation between steps instead of Concanavalin A magnetic beads. After integration, we used a Takara ICELL8 nano-dispensing system to aliquot single cells into nanowells of a 5184 well chip, identifying the nanowells that contained one and only one cell by imaging the chip. We then performed PCR enrichment of libraries in each passing nanowell using two indexed primers, and finally pooled all enriched libraries from the chip for Illumina deep sequencing to high redundancy to assess the sampling and coverage in each cell (Figure 3.11a). Libraries from each well are distinguished by unique combinations of the two indices.

The aggregate of single cell chromatin profiling closely matched profiles generated in bulk samples (Figure 3.10b), with high correlations (Pearson's r = 0.89, Figure 3.11b). Individual cells were ranked by the genome-wide number of reads, and the unique fragments are displayed in tracks for each cell. Strikingly, the majority of reads from individual cells fall within H3K27me3 blocks defined in bulk profiling, indicating high recovery in single cell chromatin profiling (Figure 3.10b). A second replicate of H3K27me3 scCUT&Tag demonstrated the reproducibility of single cell profiling. Similarly, single cell profiling of the H3K4me2 modification recapitulates genomic landscapes of accessible and active chromatin (Figure 3.10c). A significant fraction of reads in single cells fall within defined active and silenced chromatin features (Figure 3.10d, e).

The breadth of chromatin features—from ~5 nucleosomes for H3K4me2 to hundreds in H3K27me3 domains—assists the detection of chromatin features even with sparse sampling from individual cells. To assess if chromatin features in individual cells could be used to distinguish cell types, we performed scCUT&Tag to the H3K27me3 modification in H1 cells. Again, we found that a high fraction of reads fell within domains defined by bulk profiling (Figure 3.10e), with high correlations between bulk and aggregated single cell data (Pearson's r = 0.85, Figure 3.11b). Comparing a 2 Mb region encompassing the HoxB domain reveals clear histone methylation in single cell tracks specifically in H1 cells, while this region is depleted in K562 cells (Figure 3.10f). These genome-wide patterns are sufficient to discriminate single H1 cells from K562 cells with high efficiency (Figure 3.11c, d). The small fraction of

K562 cells that are mis-called have the sparsest read coverage. Thus, chromatin profiling provides a method to discriminate single cell types.

Discussion

Chromatin profiling by CUT&Tag efficiently reveals regulatory information in genomes. In contrast to RNA-seq (Svensson et al., 2018), which only measures expressed genes, chromatin profiling has the unique advantage of identifying silenced regions, which is a key aspect of establishing cell fates in development. Although methods like ATAC-seq map accessible and factor-bound sites (Buenrostro et al., 2013), the specific chromatin proteins bound at these sites must be inferred from motif or chromatin profiling data. While ChIP-based methods have been extensively used in model cell line systems, the vagaries of crosslinking and fragmenting chromatin have limited chromatin profiling by ChIP-seq to an artisan technique where each experiment requires optimization. Likewise, a recently described alternative cross-linked chromatin profiling method, ChIL-seq (Harada et al., 2018), requires many more steps than CUT&Tag and requires 3–4 days to perform all of the steps. In contrast, the CUT&Tag procedure, like CUT&RUN, is an unfixed in situ method, and is easily implemented in a standardized approach. This, combined with the cost-effectiveness of CUT&Tag, makes it appropriate for a high-throughput pipeline that can be implemented in a core facility (Janssens, 2018). It is conceivable that diverse users may provide their mixture of cells and antibody and receive processed deep sequencing files in just days. Since the first step in high-throughput CUT&Tag is antibody incubation at 4 °C, samples can be accumulated overnight in a facility and then loaded together onto a 96-well plate for robotic handling, as we previously demonstrated for AutoCUT&RUN18. With efficient use of reagents and better signal-to-noise, CUT&Tag requires even fewer reads per sample than AutoCUT&RUN, which is already much cheaper than commercial exome sequencing. While the ease and low cost of this pipeline is appealing, the primary virtue of automated chromatin profiling is the minimization of batch and handling effects, and thus maximum reproducibility. Such aspects are critical for clinical assays and testing for chromatin-targeting drugs.

We have shown that CUT&Tag provides high-quality single-cell profiles using the ICELL8 nano-dispensation system (Mezger. et al., 2018), which allows for imaging prior to reagent addition and PCR. Likewise, CUT&Tag should be suitable for the 10× Genomics encapsulation system13 by adaptation of their recently announced ATAC-seq single-cell protocol (*10xgenomics. Https://Www.10xgenomics.Com/Solutions/Single-Cell-Atac.*, n.d.). Adaptability to high-throughput single-cell platforms is possible for CUT&Tag because adapters are added in bulk, whereas previous single-cell adaptations of antibody-based profiling methods, including ChIP-seq (Rotem, 2015), ChIL-seq (Harada et al., 2018), and CUT&RUN (Hainer et al., 2018) require reactions to be performed after cells are separated. The distinct distributions of low-level untargeted accessible DNA sites and high-level CTCF-bound sites in CUT&Tag datasets suggests that by modeling the two expected underlying distributions, true binding sites can be distinguished from accessible DNA sites without using other data. An advantage

of this strategy is that the statistical distinction between true binding sites and accessible features allows characterization of two chromatin features in the same experiment, where accessible DNA sites can be annotated as well as binding sites for the targeted factor. This parsing out of the low-level ATAC-seq background from the strong targeted CUT&Tag signal makes possible de novo “multi-OMIC” CUT&Tag (Rhee & Pugh, 2011; Zentner & Henikoff, 2014). In the future, we expect that barcoding of adapters (Amini, 2014) will allow for multiple epitopes to be simultaneously profiled in single cells in large numbers, maximizing the utility of single-cell epigenomic profiling for studies of development and disease.

Methods

Biological materials

Human K562 cells were purchased from ATCC (Manassas, VA, Catalog #CCL-243) and cultured following the supplier’s protocol. H1 ES cells were obtained from WiCell (Cat#WA01-lot#WB35186). We used the following antibodies: Guinea Pig anti-Rabbit IgG (Heavy & Light Chain) antibody (Antibodies-Online ABIN101961). H3K27me3 (Cell Signaling Technology, 9733, Lot 14), H3K27ac (Millipore, MABE647), H3K4me1 (Abcam, ab8895), H3K4me2 (Upstate 07–030, Lot 26335), H3K4me3 (Active Motif, 39159), PolSer2P, PolSer5P, PolSer2+5P, PolSer7P (Cell Signaling Technology, Rpb1 CTD Antibody Sampler Kit, 54020), CTCF (Millipore 07–729), NPAT (Thermo Fisher Scientific, PA5–66839 ALX-215-065-1), and Sox2 (Abcam, ab92494).

Transposome preparation

Using the pTXB1-Tn515 expression vector, sequences downstream of lac operator were replaced with an efficient ribosome binding site, three tandem FLAG epitope tags and two IgG binding domains of staphylococcal protein A, which were PCR amplified from the pK19pA-MN vector11. The C-terminus of Protein A was separated from the transposase by a 26 residue flexible linker peptide composed of DDDKEF(GGGGS) (Skene & Henikoff, 2017). The pTXB1-Tn5 plasmid was a gift from Rickard Sandberg (Addgene plasmid # 60240) and the pK19pA-MN plasmid was a gift from Ulrich Laemmli (available through Addgene, plasmid # 86973). The 3XFlag-pA-Tn5-Fl plasmid (Addgene plasmid # 124601) was transformed into C3013 cells (NEB) following the manufacturer’s protocol. Each colony tested was inoculated into 3 mL LB medium and growth was continued at 37 °C for 4 h. That culture was used to start a 400 mL culture in 100 µg/mL carbenicillin containing LB medium (as it is more stable than ampicillin) and incubated on a shaker until it reached O.D. ~0.6, whereupon it was chilled on ice for 30 min. Fresh IPTG was added to 0.25 mM to induce expression, and the culture was incubated at 18 °C on a shaker overnight. The culture was collected by centrifugation at 10,000 rpm, 4 °C for 30 min. The bacterial pellet

was frozen in a dry ice-ethanol bath and stored at −70 °C. Protein purification was performed as previously described15 with minor modifications. Briefly, a frozen pellet was resuspended in 40 mL chilled HEGX Buffer (20 mM HEPES-KOH at pH 7.2, 0.8 M NaCl, 1 mM EDTA, 10% glycerol, 0.2% Triton X-100) including 1× Roche Complete EDTA-free protease inhibitor tablets. The lysate was sonicated 10 times for 45 s at a 50% duty cycle with output level 7 while keeping the sample chilled and holding on ice between cycles. The sonicated lysate was centrifuged at 10,000 rpm in a Fiberlite rotor at 4 °C for 30 min. A 2.5 mL aliquot of chitin slurry resin (NEB, S6651S) was packed into each of two disposable columns (Bio-rad 7321010). Columns were washed with 20 mL of HEGX Buffer. The soluble fraction was added to the chitin resin slowly, then incubated on a rotator at 4 °C overnight. The unbound soluble fraction was drained and the columns were rinsed with 20 mL HEGX and washed thoroughly with 20 mL HEGX containing Roche Complete EDTA-free protease inhibitor tablets. The chitin slurry was transferred to a 15 mL conical tube and resuspended in elution buffer (10 mL HEGX with 100 mM DTT). The tube was placed on rotator at 4 °C for ~48 h. The eluate was collected and dialyzed twice in 800 mL 2X Tn5 Dialysis Buffer (100 mM HEPES-KOH pH 7.2, 0.2 M NaCl, 0.2 mM EDTA, 2 mM DTT, 0.2% Triton X-100, 20% Glycerol). The dialyzed protein solution was concentrated using an Amicon Ultra-4 Centrifugal Filter Units 30 K (Millipore UFC803024), and sterile glycerol was added to make a final 50% glycerol stock of the purified protein.

To generate the pA-Tn5 adapter transposome, 16 µL of a 100 µM equimolar mixture of preannealed Tn5MEDS-A and Tn5MEDS-B oligonucleotides15 were mixed with 100 µL of 5.5 µM pA-Tn5 fusion protein. The mixture was incubated on a rotating platform for 1 h at room temperature and then stored at −20 °C. The complex is stable at room temperature, with no detectable loss of potency after 10 days on the benchtop (Figure 3.12a), and without noticeable loss of data quality (Figure 3.12b). Unexpectedly, this extended room temperature incubation resulted in a 1–2 order-of-magnitude increase in the number of tagmented E. coli fragments (Figure 3.12c), which can be used as a calibration standard within a series using a constant amount of pA-Tn5. This observation suggests that the E. coli DNA that co-purifies with pA-Tn5 is subject to tagmentation both during room temperature incubation and during

tagmentation in situ, where the dramatic increase seen with pre-incubation results from subsequent trapping of tagmented pA-Tn5-bound DNA within the cell. In support of this interpretation, we note that E. coli carry-over DNA suitable for calibration is also released by pA-MNase in CUT&RUN reactions during digestion32. A likely explanation for the trapping of these different fusion protein-bound DNAs within cells is that the protein-protein interaction domains of Protein A that are specific for IgG bind non-specifically to cellular proteins, whereupon addition of divalent cation results in MNase digestion and release (pA-MNase) or tagmentation (pA-Tn5).

CUT&Tag for bench-top application

Cells were harvested, counted and centrifuged for 3 min at 600×g at room temperature. Aliquots of cells (60–500,000 cells), were washed twice in 1.5 mL Wash Buffer (20 mM HEPES pH 7.5; 150 mM NaCl; 0.5 mM Spermidine; 1× Protease inhibitor cocktail) by gentle pipetting. Concanavalin A coated magnetic beads (Bangs Laboratories) were prepared as described9 and 10 µL of activated beads were added per sample and incubated at RT for 15 min. We observed that binding cells to beads at this step increases binding efficiency. The unbound supernatant was removed and bead-bound cells were resuspended in 50–100 µL Dig-wash Buffer (20 mM HEPES pH 7.5; 150 mM NaCl; 0.5 mM Spermidine; 1× Protease inhibitor cocktail; 0.05% Digitonin) containing 2 mM EDTA and a 1:50 dilution of the appropriate primary antibody. Primary antibody incubation was performed on a rotating platform for 2 h at room temperature (RT) or overnight at 4 °C. The primary antibody was removed by placing the tube on the magnet stand to clear and pulling off all of the liquid. To increase the number of Protein A binding sites for each bound antibody, an appropriate secondary antibody (such as Guinea Pig anti-Rabbit IgG antibody for a rabbit primary antibody) was diluted 1:50 in 50–100 µL of Dig-Wash buffer and cells were incubated at RT for 30 min. Cells were washed using the magnet stand 2–3× for 5 min in 0.2–1 mL Dig-Wash buffer to remove unbound antibodies. A 1:200 dilution of pA-Tn5 adapter complex (~0.04 µM) was prepared in Dig-med Buffer (0.05% Digitonin, 20 mM HEPES, pH 7.5, 300 mM NaCl, 0.5 mM Spermidine, 1× Protease inhibitor cocktail). After removing the liquid on the magnet stand, 50–100 µL was added to the cells with gentle vortexing, which was incubated with pA-Tn5 at RT for 1 h. Cells were washed 2–3× for 5 min in 0.2–1 mL Dig-med Buffer to remove unbound pA-Tn5 protein. Next, cells were resuspended in 50–100 µL Tagmentation buffer (10 mM $MgCl_2$ in Dig-med Buffer) and incubated at 37 °C for 1 h. To stop tagmentation, 2.25 µL of 0.5 M EDTA, 2.75 µL of 10% SDS and 0.5 µL of 20 mg/mL Proteinase K was added to 50 µL of sample, which was incubated at 55 °C for 30 min or overnight at 37 °C, and then at 70 °C for 20 min to inactivate Proteinase K. To extract the DNA, 122 µL Ampure XP beads were added to each tube with vortexing, quickly spun and held 5 min. Tubes were placed on a magnet stand to clear, then the liquid was carefully withdrawn. Without disturbing the beads, beads were washed twice in 1 mL

80% ethanol. After allowing to dry ~5 min, 30–40 µL of 10 mM Tris pH 8 was added, the tubes were vortexed, quickly spun and allowed to sit for 5 min. Tubes were placed on a magnet stand and the liquid was withdrawn to a fresh tube.

To amplify libraries, 21 µL DNA was mixed with 2 µL of a universal i5 and a uniquely barcoded i7 primer33, using a different barcode for each sample. A volume of 25 µL NEBNext HiFi 2× PCR Master mix was added and mixed. The sample was placed in a Thermocycler with a heated lid using the following cycling conditions: 72 °C for 5 min (gap filling); 98 °C for 30 s; 14 cycles of 98 °C for 10 s and 63 °C for 30 s; final extension at 72 °C for 1 min and hold at 8 °C. Post-PCR clean-up was performed by adding 1.1× volume of Ampure XP beads (Beckman Counter), and libraries were incubated with beads for 15 min at RT, washed twice gently in 80% ethanol, and eluted in 30 µL 10 mM Tris pH 8.0.

A detailed, step-by-step protocol can be found at https://www.protocols.io/view/bench-top-cut-amp-tag-wnufdew/abstract

High-throughput CUT&Tag

For high-throughput 96-well microplate application, cells were first permeabilized and incubated with the primary antibodies before binding to beads. Two biological replicates of human K562 and H1 ES cells were washed twice with Wash Buffer, resuspended in Dig-wash buffer with 2 mM EDTA and arrayed in a 96-well plate. Permeabilization before antibody incubation varied from 1 to 5 h. Then, dilutions of appropriate antibodies were added (making final antibody concentrations 1:50) as duplicates. Cells were incubated with primary antibodies overnight. The next day, 10 µL of activated Concanavalin A coated magnetic beads were added to each sample, mixed gently and incubated at room temperature for 10 min. The plate was placed on a magnetic plate holder and supernatants were discarded. Appropriate secondary antibodies were prepared as 1:50 dilutions in Dig-wash and added to each well. Cells were washed three times with Dig-wash and then incubated with 1:200 dilution of pA-Tn5 adapter complex in Dig-med buffer at RT for 1 h. Cells were washed 3× for 5 min in Dig-med Buffer and resuspended in 50 µL Tagmentation buffer and incubated at 37 °C for 1 h. To stop tagmentation, 2.25 µL of 0.5 M EDTA, 2.75 µL of 10% SDS and 0.5 µL of 20 mg/mL Proteinase K was added to the sample, which was incubated at 55 °C for 30 min and then at 70 °C for 20 min to inactivate Proteinase K. Samples were held at 4 °C overnight until ready to continue. A 1.1× volume of AMPure XP beads was added to each well, vortexed and incubated at room temperature for 10–15 min. The plate was placed on a magnet and unbound liquid was removed. Beads were gently rinsed twice with 80% ethanol, and DNA was eluted with 35 µL of 10 mM Tris-HCl pH 8. 30 µL of eluted DNA was amplified by PCR as described above.

DNA sequencing and data processing

The size distribution of libraries was determined by Agilent 4200 TapeStation analysis, and libraries were mixed to achieve equal representation as desired aiming for a final concentration as recommended by the manufacturer. Paired-end Illumina sequencing was performed on the barcoded

libraries following the manufacturer's instructions. Paired-end reads were aligned using Bowtie2 version 2.2.5 with options: –local–very-sensitive-local–no-unal–no-mixed–no-discordant–phred33 -I 10 -X 700. Because of the very low background with CUT&Tag, typically 3 million paired-end reads suffice for nucleosome modifications, even for the human genome. For maximum economy, up to 96 barcoded samples per 2-lane flow cell can be pooled for 25 × 25 bp sequencing. For peak calling, parameters used were macs2 callpeak—t input_file –p 1e-5 –f BEDPE/BED(Paired End vs. Single End sequencing data) –keep-dup all –n out_name.

Single-cell CUT&Tag

Approximately 50,000 exponentially growing K562 cells were processed by centrifugation between buffer and reagent exchanges in low-retention tubes throughout. Centrifugations were performed at 600×g for 3 min in a swinging bucket rotor for the initial wash and incubation steps, and then at 300×g for 3 min after pA-Tn5 binding. Cells were collected and washed twice with 1 mL Wash Buffer (20 mM HEPES, pH 7.5; 150 mM NaCl; 0.5 mM Spermidine, 1× Protease inhibitor cocktail) at room temperature. Nuclei were isolated by permeabilizing cells in NP40-Digitonin Wash Buffer (0.01% NP40, 0.01% Digitonin in wash buffer) and resuspended in 1 mL of NP40-Digitonin Wash buffer with 2 mM EDTA. Antibody was added at a 1:50 dilution and incubated on a rotator at 4 °C overnight. Permeabilized cells were then rinsed once with NP40-Digitonin Wash buffer and incubated with anti-Rabbit IgG antibody (1:50 dilution) in 1 mL of NP40-Digitonin Wash buffer on a rotator at room temperature for 30 min. Nuclei were then washed 3× for 5 min in 1 mL NP40-Digitonin Wash buffer to remove unbound antibodies. For pA-Tn5 binding, a 1:100 dilution of pA-Tn5 adapter complex was prepared in 1 mL NP40-Dig-med-buffer (0.01% NP40, 0.01% Digitonin, 20 mM HEPES, pH 7.5, 300 mM NaCl, 0.5 mM Spermidine, 1× Protease inhibitor cocktail), and permeabilized cells were incubated with the pA-Tn5 adapter complex on a rotator at RT for 1 h. Cells were washed 3× for 5 min in 1 mL NP40-Dig-med-buffer to remove excess pA-Tn5 protein. Cells were resuspended in 150 µL Tagmentation buffer (10 mM MgCl2 in NP40-Dig-med-buffer) and incubated at 37 °C for 1 h. Tagmentation was stopped by adding 50 µL of 4× Stop Buffer (40.4 mM EDTA and 2 mg/mL DAPI) and the sample was held on ice for 30 min.

The SMARTer ICELL8 single-cell system (Takara Bio USA, Cat. #640000) was used to array single cells as described for scATAC-seq12. DAPI-stained nuclei were visualized under the microscope and if there were clumps, they were strained through 10 micron cell strainers. Cells were counted using a hematocytometer and diluted at 28 cells/µL in 0.5× PBS and 1× Second Diluent (Takara Bio USA, Cat. # 640196). Cells were loaded to a source loading plate. Control wells containing 0.5× PBS (25 µL) and fiducial mix (25 µL) (Takara Bio USA, Cat. #640196) were also included in the source loading plate. Using the ICELL8 MultiSample NanoDispenser (MSND) FLA program, cells were dispensed into a SMARTer ICELL8 350 v chip (Takara Bio USA, Cat. # 640019) at 35 nanoliter per well. After cell dispense was complete, chips were sealed with the imaging film (Takara Bio USA, Cat. #640109) and centrifuged at 400×g for 5 min at room temperature and imaged using the ICELL8 imaging station (Takara Bio USA).

Images were analyzed using automated microscopy image analysis software (CellSelect, Takara Bio USA). Since cells were stained only with DAPI, they were propidium iodide negative, so that permeabilized cells would not be excluded by default software settings. Additional single cells were manually selected for dispensing using a manual triaging procedure. Immediately following imaging, the filter file, which notes single-cell containing wells and control wells, was generated. We typically obtained ~1000 single cells per chip. All of the following reagents were added to the selected set of wells which contained single cells. To index the whole chip, 72 i5 and 72 i7 unique indices33 were dispensed at 7.32 μM using ICELL8 MSND FLA program using the index 1 and index 2 filtered dispense tool respectively at 35 nanoliter per well. NEBNext High-Fidelity 2X PCR Master Mix (NEB, M0541L) was dispensed twice using the ICELL8 MSND Single Cell/TCR program for the filtered dispense tool at 50 nL per well. The chip was sealed and centrifuged at 2250×g at 4 °C for at least 5 min after each dispense. The chip was sealed with a TE Sealing film (Takara Bio USA, Cat. #640109) and on-chip PCR was performed using a SMARTer ICELL8 Thermal Cycler (Takara Bio USA) as follows: 5 min at 72 °C and 2 min at 98 °C followed by 15 cycles of 10 s at 98 °C, 30 s at 60 °C, and 5 s at 72 °C, with a final extension at 72 °C for 1 min. PCR products were collected by centrifugation at ~2250×g for 20 min using the supplied SMARTer ICELL8 Collection Kit (Takara Bio USA, Cat.#640048).

Pooled libraries were purified using Ampure XP beads (Beckman Counter) in a 1:1.1 ratio. Briefly, libraries were incubated with beads for 15 min at RT, washed twice in 80% ethanol, and eluted in 10 mM Tris pH 8.0. Paired-end 25 × 8 × 8 × 25 bp Illumina sequencing was performed on the pooled barcoded libraries following the manufacturer's instructions. Paired-end reads were aligned using Bowtie2 version 2.2.5 with options:–local–very-sensitive-local–no-unal–no-mixed–no-discordant–phred33 -I 10 -X 700.

Data availability

Publically available datasets analyzed in this work are available in Supplementary Note 1. All sequencing data generated in this study have been deposited in GEO under accession GSE124557. The 3XFlag-pA-Tn5-Fl plasmid has been deposited with Addgene (#124601). Source data for the figures can be found in the Source Data file. All other data are available from the authors upon reasonable request.

Author Contribution

H.S.K. and S.H. designed and performed all experiments. C.A.C and E.S.P. and T.D.B. assisted with the experiments. H.S.K., S.J.W., J.G.H., K.A., and S.H. developed algorithms and analyzed the data. H.S.K, K.A., and S.H. wrote the manuscript.

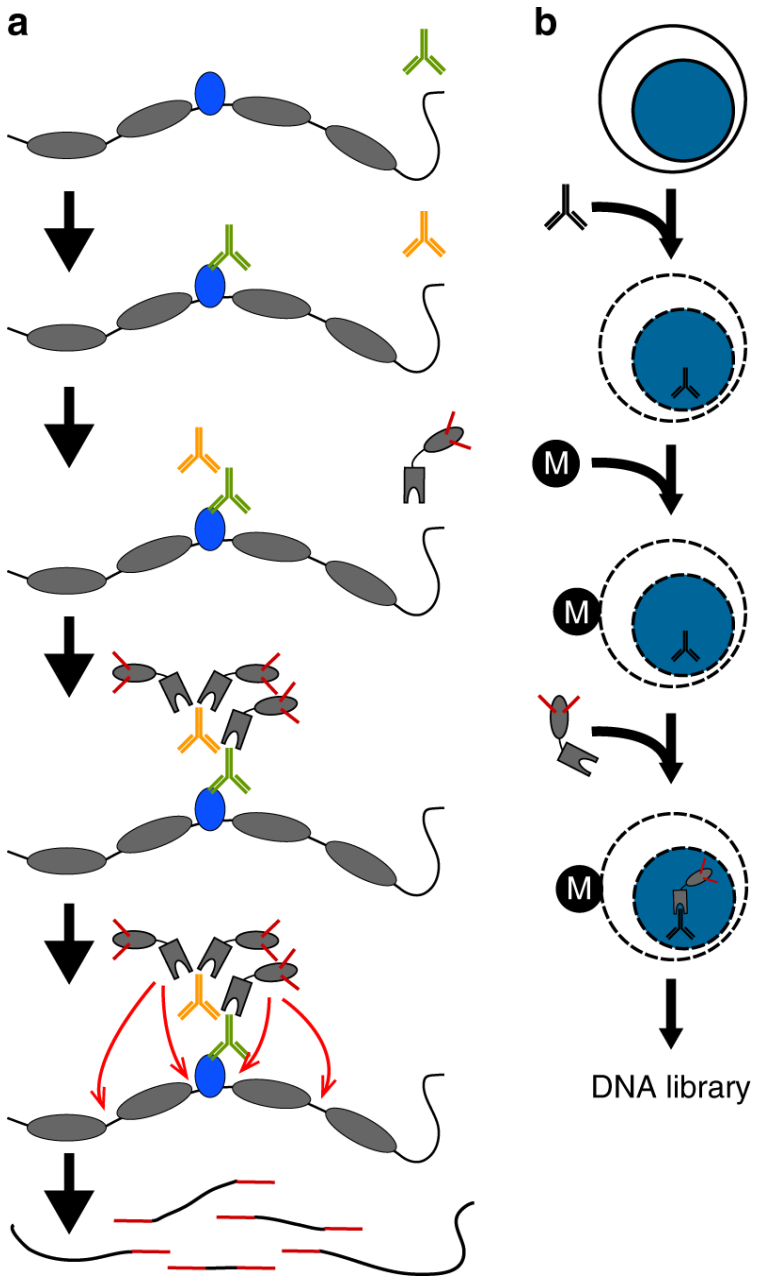


Figure 3.1: In situ tethering for CUT&Tag chromatin profiling.
a) The steps in CUT&Tag. Added antibody (green) binds to the target chromatin protein (blue) between nucleosomes (gray ovals) in the genome, and the excess is washed away. A second antibody (orange) is added and enhances tethering of pA-Tn5 transposome (gray boxes) at antibody-bound sites. After washing away excess transposome, addition of Mg++ activates the transposome and integrates adapters (red) at chromatin protein binding sites. After DNA purification genomic fragments with adapters at both ends are enriched by PCR. b) CUT&Tag is performed on a solid support. Unfixed cells or nuclei (blue) are permeabilized and mixed with antibody to a target chromatin protein. After addition and binding of cells to Concanavilin A-coated magnetic beads (M), all further steps are performed in the same reaction tube with magnetic capture between washes and incubations, including pA-Tn5 tethering, integration, and DNA purification.

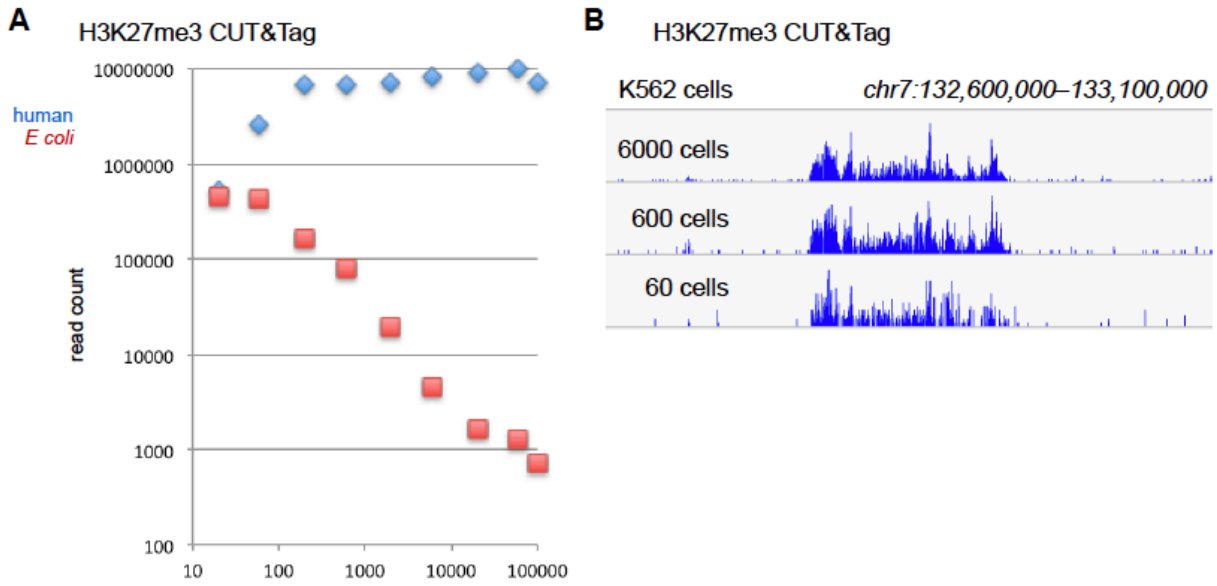


Figure 3.2: E. coli carry-through DNA provides a spike-in proxy for CUT&Tag.
a) Mapping of paired-end reads from a CUT&Tag human cell-dilution experiment to the human and Escherichia coli genomes shows that the number of E. coli reads increases proportionally with reduced cell numbers over at least 3 orders of magnitude. This indicates that calibration between samples in a treatment series can be achieved for CUT&Tag by dividing the number of reads mapping to the experimental genome by the number of E. coli reads, obviating the need for a heterologous spike-in standard. b) Chromatin landscapes of the H3K27me3 histone modification in K562 cells from limited cell numbers.

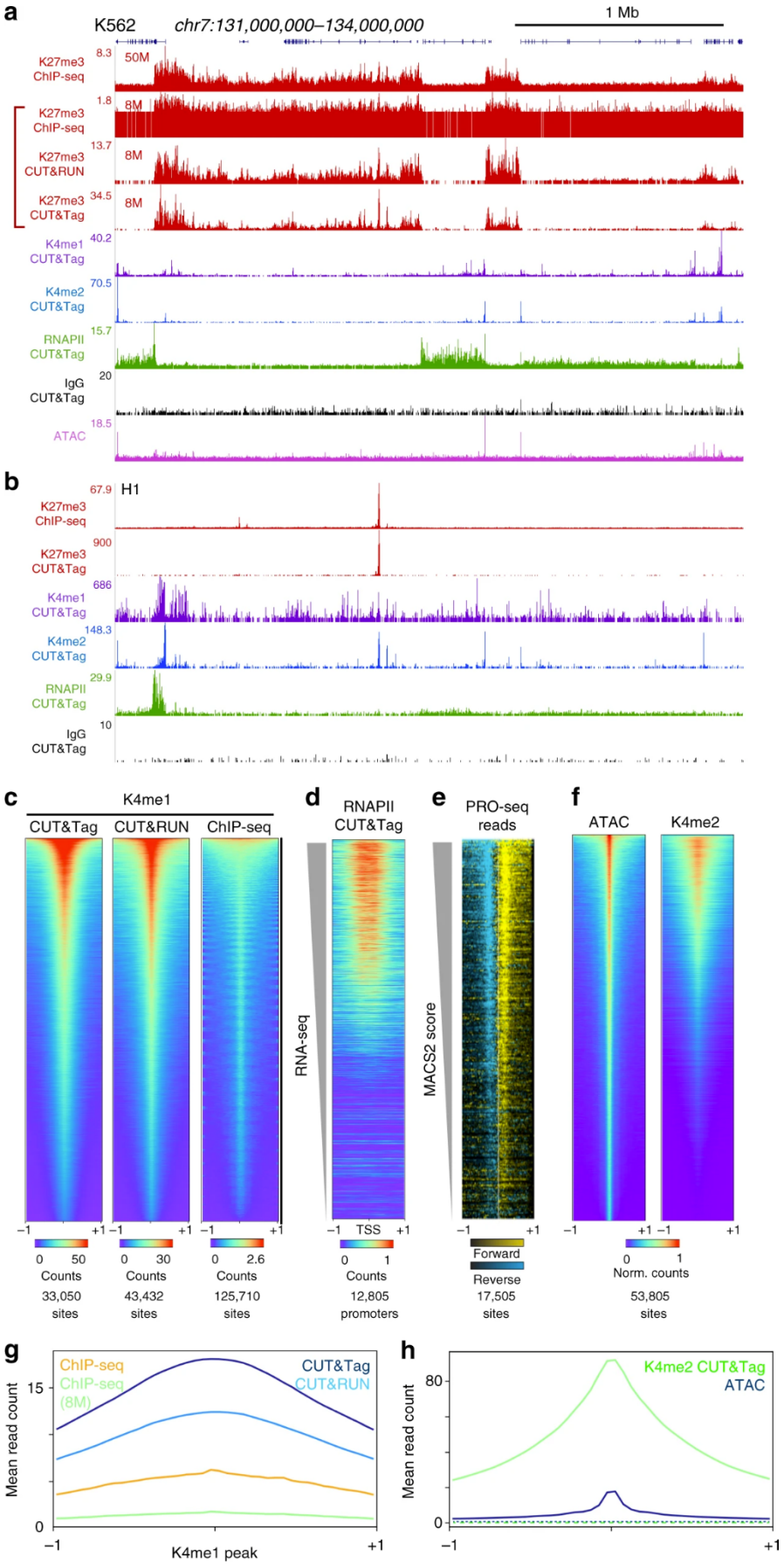

a
K562
chr7:131,000,000–134,000,000
1 Mb
K27me3 ChIP-seq
K27me3 ChIP-seq
K27me3 CUT&RUN
K27me3 CUT&Tag
K4me1 CUT&Tag
K4me2 CUT&Tag
RNAPII CUT&Tag
IgG CUT&Tag
ATAC
b
H1
c
K4me1
CUT&Tag
CUT&RUN
ChIP-seq
Counts
33,050 sites
43,432 sites
125,710 sites
d
RNAPII CUT&Tag
RNA-seq
TSS
12,805 promoters
e
PRO-seq reads
MACS2 score
Forward
Reverse
17,505 sites
f
ATAC
K4me2
Norm. counts
53,805 sites
g
ChIP-seq
ChIP-seq (8M)
CUT&Tag
CUT&RUN
Mean read count
K4me1 peak
h
K4me2 CUT&Tag
ATAC
Mean read count

Figure 3.3: CUT&Tag for histone modification profiling and RNAPII.
a) Representative chromatin landscapes across a 3 Mb segment of the human genome generated by the indicated method. For H3K27me3, we downsampled ChIP-seq and CUT&RUN datasets to the same total mapped read counts as CUT&Tag for direct comparison. The high background in downsampled ChIP-seq is from singleton reads distributed across the genome. b) Same as a except for H1 ES cells. c) Comparison of profiling methods for the H3K4me1 histone modification in K562 cells. The same antibody was used in all experiments. Peaks were called and ordered for each dataset using MACS2. Each dataset was downsampled to the same read depth for comparison and plotted on their called peaks. Color intensities are scaled to the maximum read count at peaks in each dataset. d) Detection of gene activity by RNAPII CUT&Tag. Gene promoters were ordered by associated RNA-seq counts (gray wedge) and read counts from RNAPII S2/5p CUT&Tag were plotted on these sites. e) Active RNAPII is enriched at RNAPII CUT&Tag peaks. Peaks were called from RNAPII S2/5p CUT&Tag and ordered using MACS2 (gray wedge). PRO-seq reads were displayed onto these positions for (+) strand reads (yellow) and (–) strand reads (blue). f) Comparison of ATAC-seq and H3K4me2 CUT&Tag profiling in K562 cells. Peaks were called on ATAC-seq data and heat maps were produced as in c. The top and bottom 2.5% of peaks were discarded to remove outliers. g) Metaplot comparison of H3K4me1 histone modification signal in CUT&RUN, CUT&Tag, and ChIP-seq in K562 cells, averaged at the top 10,000 peaks detected by MACS2 in ChIP-seq data. Profiling with the same antibody was compared at the downsampled read depths of 8 million mapped reads for all three methods (blue, cyan, and green), and for 40 million mapped reads (orange) from ChIP-seq. h) Metaplot comparison of ATAC-seq and H3K4me2 CUT&Tag profiling in K562 cells. 53,805 peaks were called on ATAC-seq data using MACS2, and read counts from each method were averaged across the intervals. The top and bottom 2.5% of peaks were discarded to remove outliers. Read counts at 17,000 randomly-chosen intervals for each dataset are displayed as dotted lines. Source data are available in the Source Data file

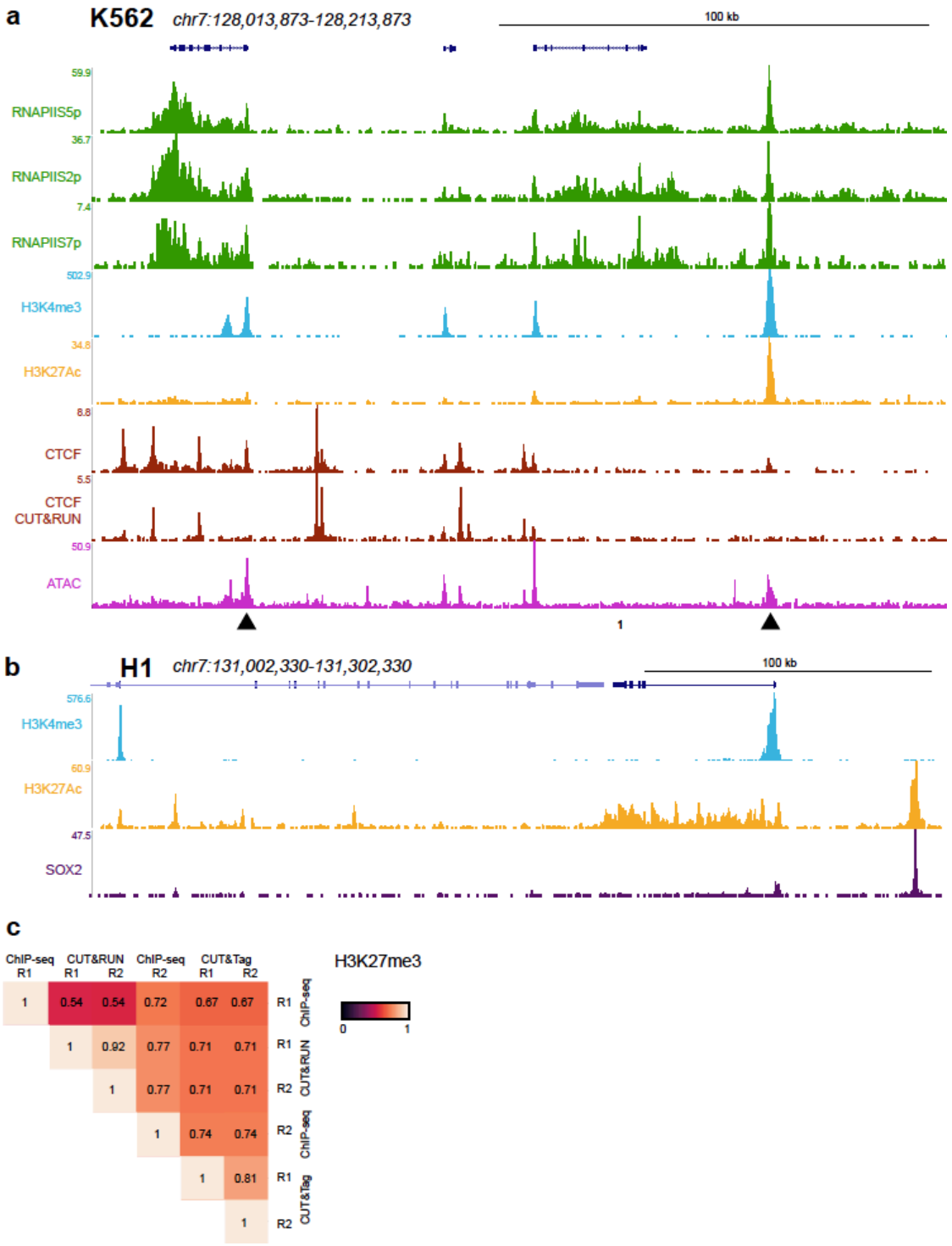


Figure 3.4: Chromatin profiling by CUT&Tag.
a-b) Additional CUT&Tag profiling landscapes for histone modifications and transcription factors in (a) K562 and (b) H1 ES cells. The arrowheads indicate hypersensitive (ATAC) sites that appear as CTCF CUT&Tag peaks but are absent from CTCF CUT&RUN profiling. c) Hierarchically clustered correlation matrix of CUT&Tag, CUT&RUN, and ChIP-seq profiling replicates (R1 and R2) for the H3K27me3 histone modification. The same antibody was used in all experiments. Pearson correlations were calculated using the log2-transformed values of read counts split into 500 bp bins across the genome.

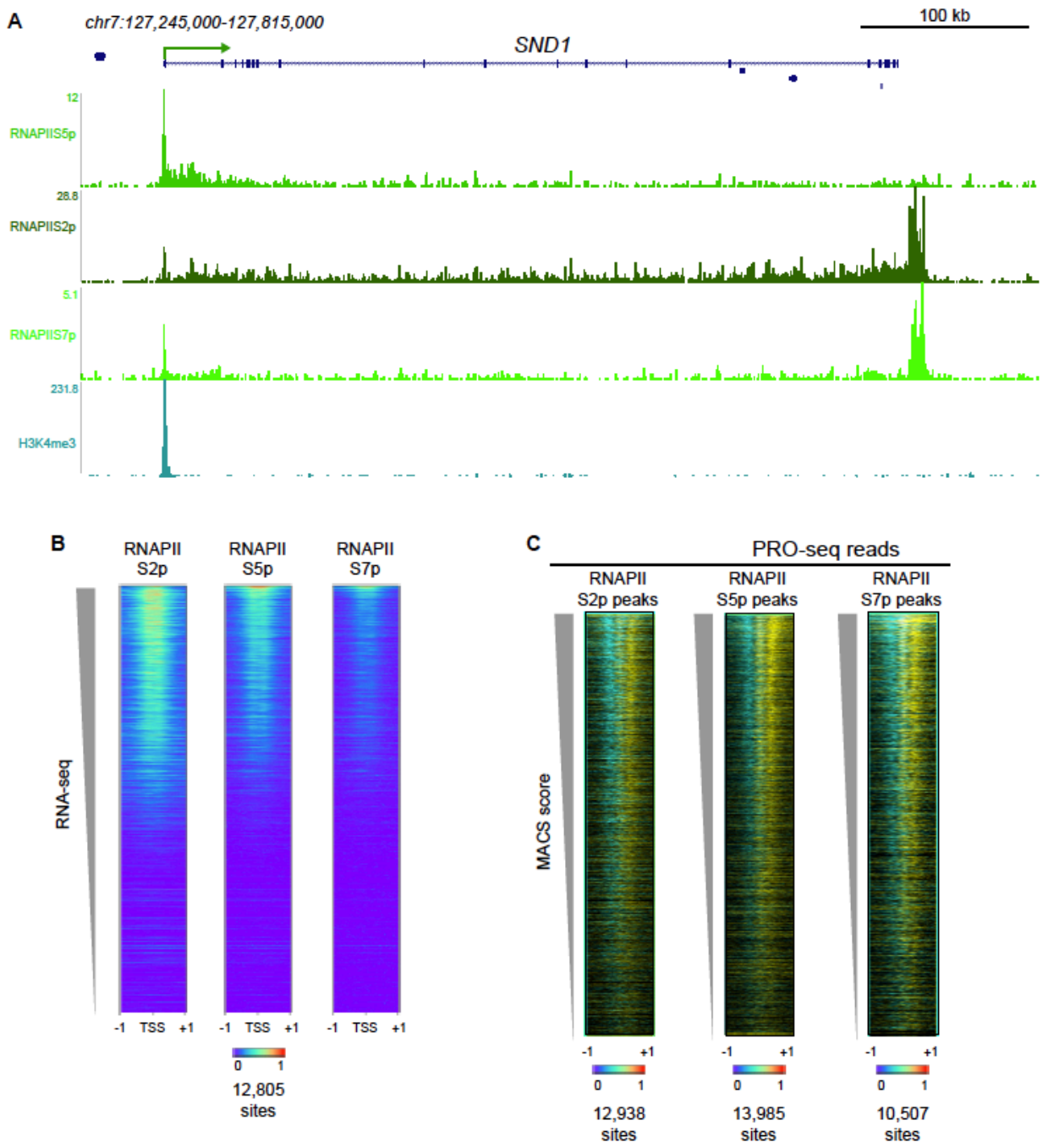


Figure 3.5: Profiling of gene activity by CUT&Tag.
a) RNAPII CUT&Tag marks the promoters of active genes. Chromatin landscapes of CUT&Tag profiling across the SND1 gene. b) RNAPII CUT&Tag on gene promoters. Annotated genes were ordered by expression as determined by RNAseq read counts, and RNAPII CUT&Tag reads were plotted for three different RNAPII modifications. c) Active RNAPII is enriched at RNAPII CUT&Tag peaks. Peaks were called from RNAPII CUT&Tag for different modifications using MACS2. PRO-seq reads were displayed onto these positions for (+) strand reads (yellow)
and (–) strand reads (blue).

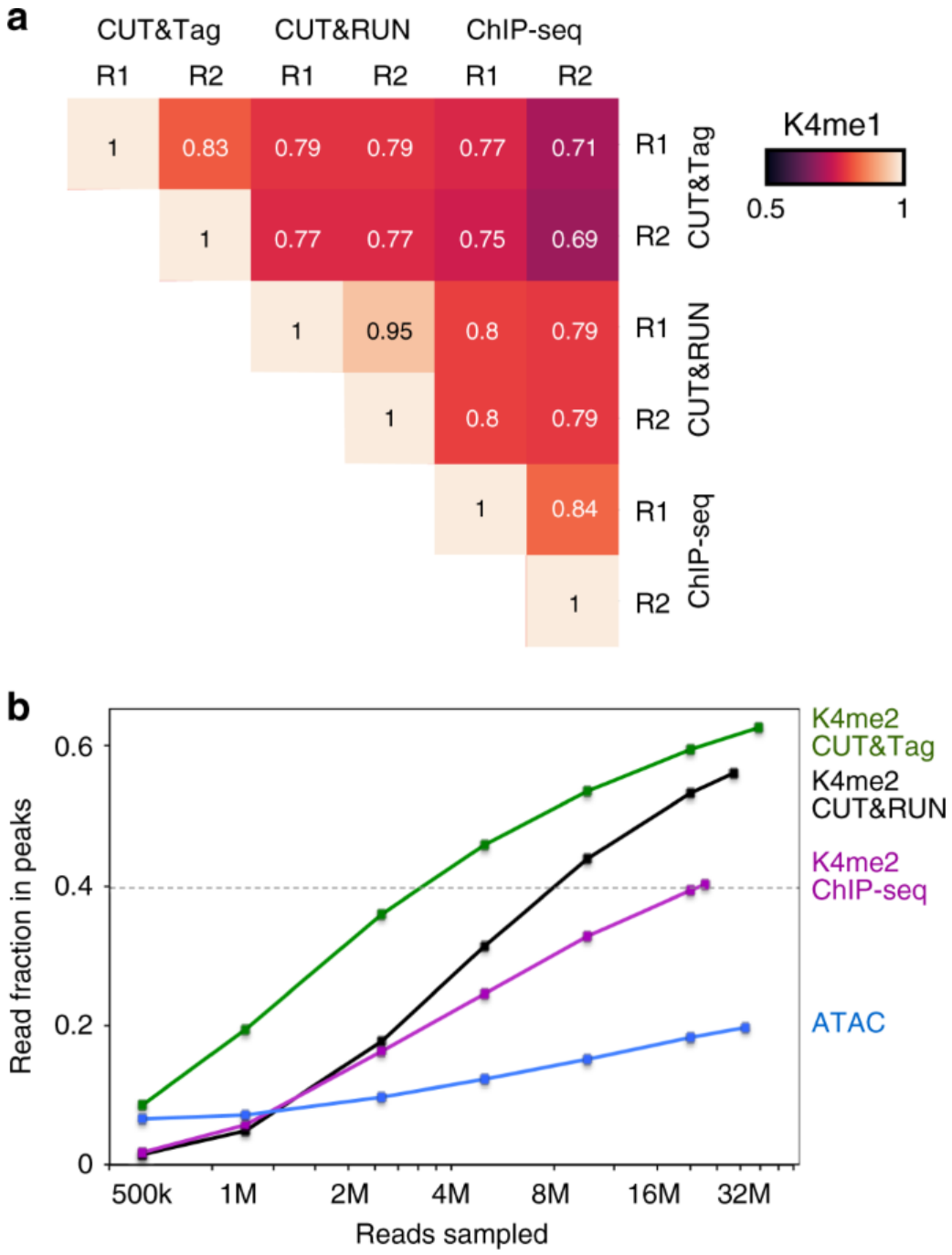


Figure 3.6: Reproducibility and efficiency of CUT&Tag.
a) Hierarchically clustered correlation matrix of CUT&Tag replicates (R1 and R2) and with CUT&RUN and ChIP-seq profiling for the H3K4me1 histone modification. The same antibody was used in all experiments. Pearson correlations were calculated using the log2-transformed values of read counts split into 500 bp bins across the genome. b) Efficiency of peak-calling between methods. Mitochondrial reads were removed from datasets from each method profiling the H3K4me2 histone modification. The remaining read counts were downsampled to varying depths, and then used to call peaks using MACS2. The summed number of reads falling within called peaks in each dataset was plotted. Source data are available in the Source Data file

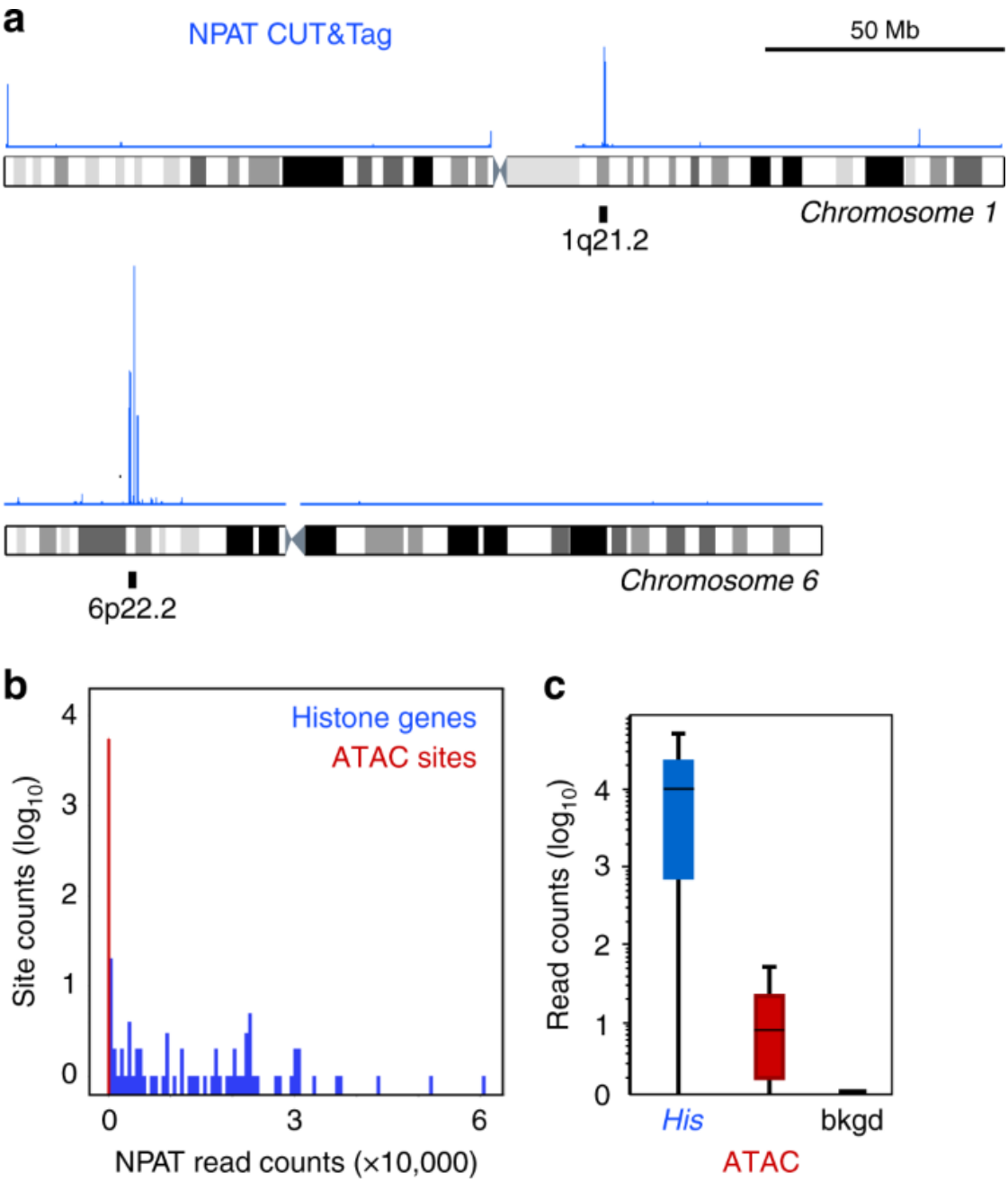


Figure 3.7: CUT&Tag profiling of the NPAT chromatin factor and chromatin accessibility.
a) Ideograms of chromosome 1 and chromosome 6 with the clusters of replication-dependent histone genes (6 genes at 1q21.2 and 55 genes at 6p22.2) indicated. An NPAT CUT&Tag profiling track is displayed over each chromosome. The major NPAT peaks fall at the histone genes. b) Distribution of read counts in CUT&Tag profiling. Called accessible sites from ATAC-seq data were segregated into those at histone genes and other ATAC sites. Read counts from NPAT CUT&Tag were plotted for each category. c) Boxplots of NPAT CUT&Tag signal at the promoters of replication-dependent histone genes (His), at other accessible sites called from ATAC-seq data, and at a random selection of genomic background sites (bkgd). Source data are available in the Source Data file

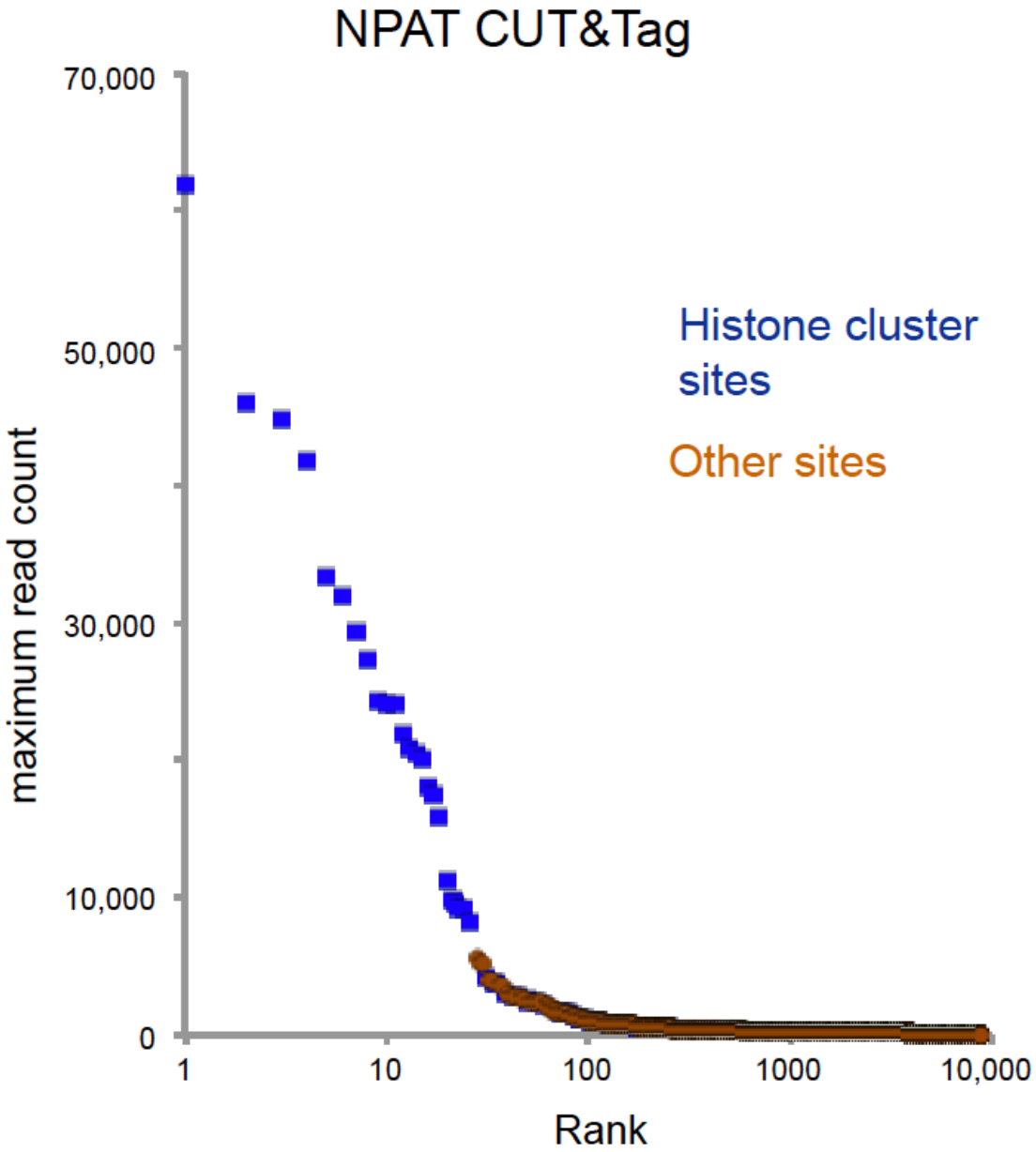


Figure 3.8: Discrimination of NPAT binding sites from accessible sites.
We called 8,689 significant peaks using MACS2 on NPAT CUT&Tag data from K562 cells. The ordered list of sites is plotted against the maximum read count at each site. Sites falling within the histone gene clusters were color-coded blue, and the remaining sites are colored orange.

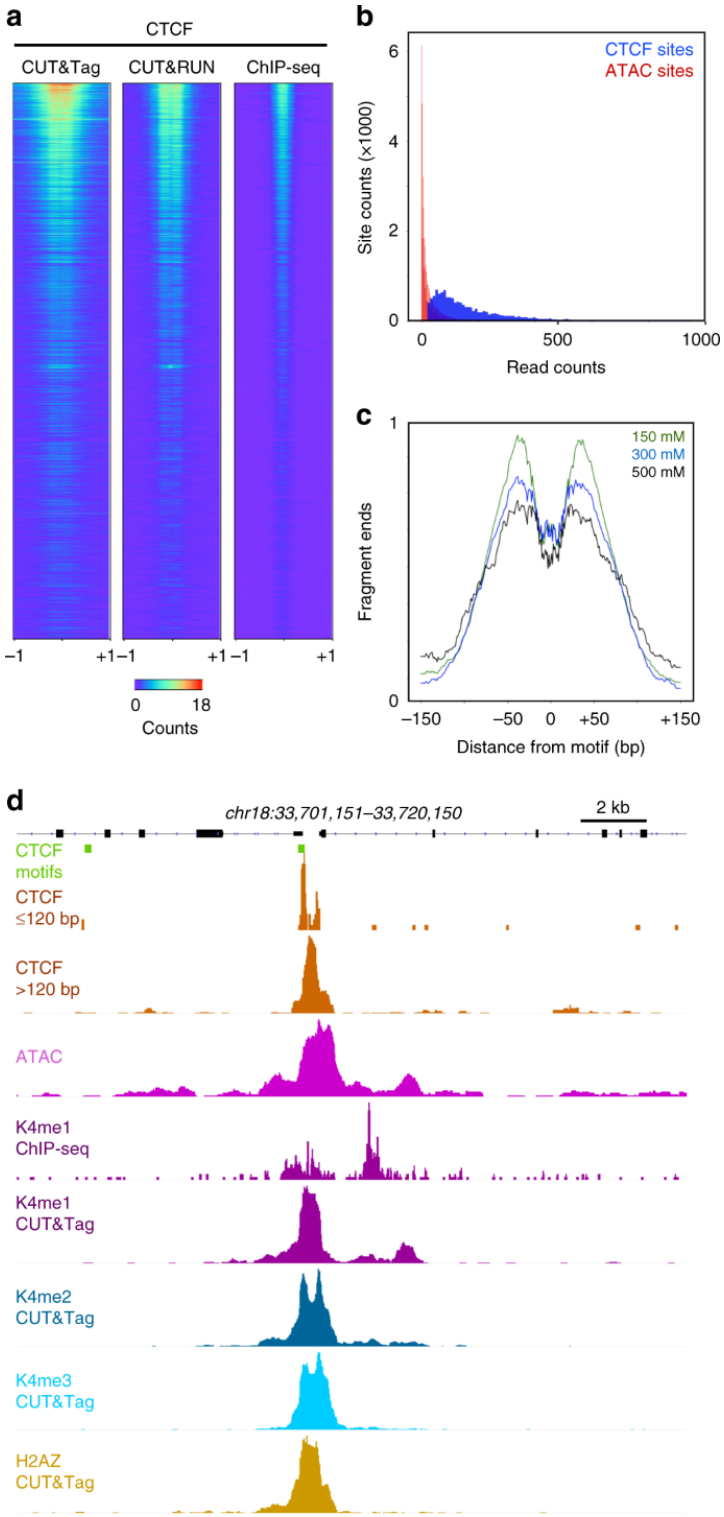


Figure 3.9: CUT&Tag profiling of the CTCF DNA-binding protein.
a) Comparison of methods for CTCF mapping. CTCF motifs in the genome were ranked by e-value, datasets from each method were downsampled to the same read depth, and then read counts were plotted on the fixed order of sites. b) Distribution of read counts in CTCF CUT&Tag profiling. Sites were called from CUT&RUN profiling (blue) and at non-overlapping accessible sites (red, called from ATAC-seq). Read counts from CTCF CUT&Tag were plotted for each category. c) Resolution of CUT&Tag.

Mean plots of fragment end positions from CTCF CUT&Tag centered over CTCF motifs in called peaks. Three different NaCl concentrations were used in wash buffers during and after pA-Tn5 tethering. Data are represented as a fraction of the maximum signal within the interval. d) Resolved structure of a CTCF binding site. The promoter of the SLC39A6 gene on chromosome 18 shows the chromatin features around a CTCF-bound site. Source data are available in the Source Data file

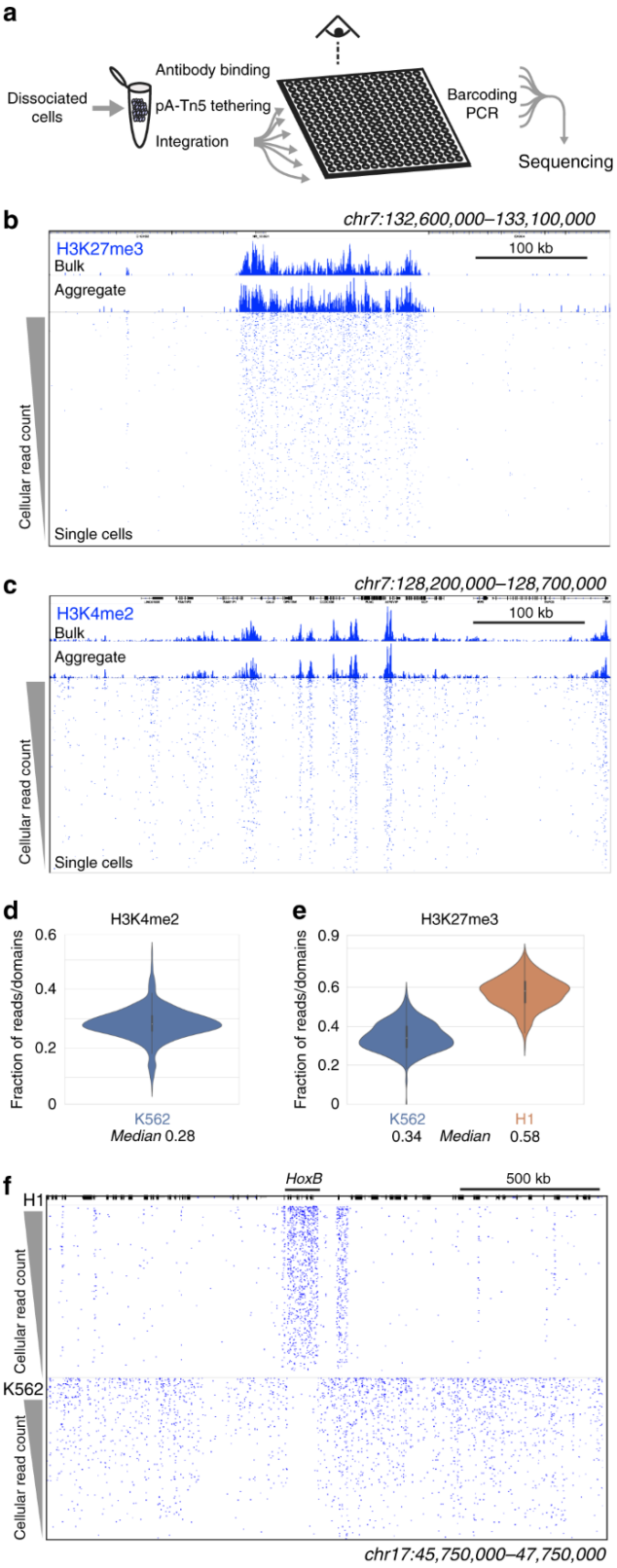


Figure 3.10: Chromatin profiling of individual cells.

a) Single cell CUT&Tag (scCUT&Tag) processing. All steps from antibody incubations through adapter tagmentation are done on a population of permeabilized unfixed cells. Individual cells are then dispensed

into nanowells of a Takara ICELL8 chip. After verifying nanowells with single cells by microscopy, combinations of two indexed barcoded primers are added to each well and fragment libraries are enriched by PCR. Libraries from the chip are pooled for multiplex sequencing. b) A chromatin landscape across a 500 kb segment of the human genome is shown for H3K27me3 CUT&Tag on K562 cells. Tracks from bulk CUT&Tag, aggregated scCUT&Tag, and for 956 single cells are shown. Single cells are ordered by total read counts in each cell. c) A chromatin landscape across a 500 kb segment of the human genome for H3K4me2 CUT&Tag on K562 cells. Tracks from bulk CUT&Tag, aggregated scCUT&Tag, and for 808 single cells are shown. Single cells are ordered by total read counts in each cell. d) Fraction of reads in single K562 cells falling within called active peaks for the H3K4me2 histone modification using stringent criteria. Narrow peaks were called using MACS2 on bulk profiling data, and reads from scCUT&Tag were assigned to those peaks. e) Fraction of reads in single K562 and H1 cells falling within called silenced domains for the H3K27me3 histone modification. Domains were called using SEACR on bulk profiling data for each cell type, and reads from scCUT&Tag were assigned to those domains. f) Comparison of chromatin landscapes in H1 and K562 single cells across a 2 Mb segment including the HoxB locus. Four hundred and seventy-nine single cells of each type were ordered by total read counts. Source data are available in the Source Data file

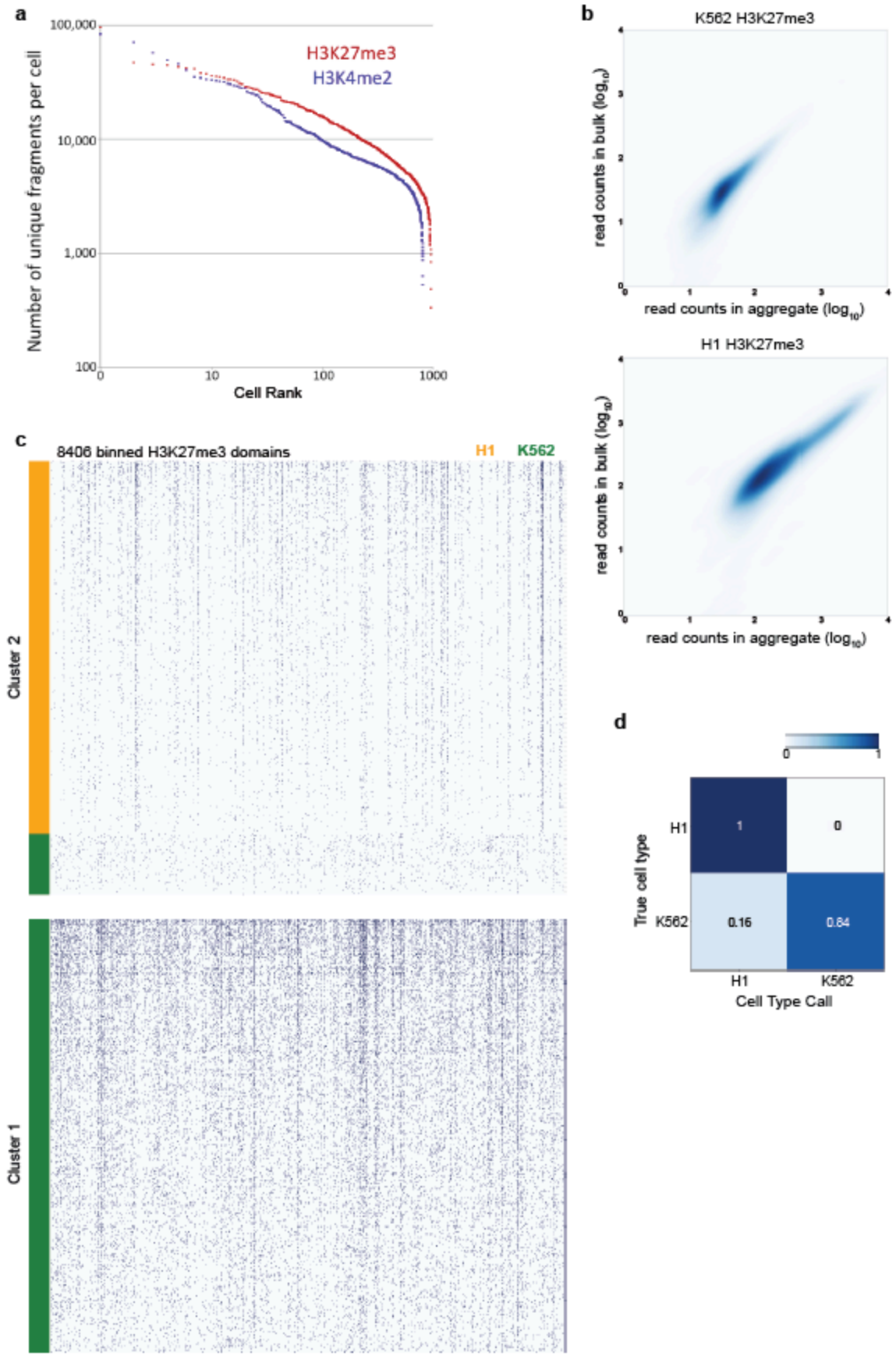

a
100,000
10,000
1,000
100
Number of unique fragments per cell
H3K27me3
H3K4me2
0
10
100
1000
Cell Rank
b
K562 H3K27me3
read counts in bulk (log10)
read counts in aggregate (log10)
H1 H3K27me3
read counts in bulk (log10)
read counts in aggregate (log10)
c
8406 binned H3K27me3 domains
H1
K562
Cluster 2
Cluster 1
d
0
1
H1
K562
True cell type
1
0
0.16
0.84
H1
K562
Cell Type Call

Figure 3.11: Single-cell CUT&Tag fragment recovery.
a) Plot of the number of unique reads from each single cell after scCUT&Tag. 956 single K562 cells in H3K27me3 scCUT&Tag and 808 single K562 cells in H3K4me2 scCUT&Tag. Cells were ranked based on their number of unique reads. b) Correlation of genomic profiles for bulk CUT&Tag and aggregated scCUT&Tag experiments with K562 and H1 cells. The R-value for K562 cells is 0.85, and that for H1 cells is 0.88. c) Clusters of single cells based on H3K27me3 profiles. Reads from 479 H1 and 479 K562 cells were binned into a combined list of H3K27me3 domains called from bulk profiles and sorted by Kmeans with k=2. Cells are color-coded on the left end of the matrix. All H1 single cells are assigned to one cluster, while 84% of K562 single cell profiles are assigned to the other cluster. The 78 mis-assigned K562 profiles have very low coverage. d) Confusion matrix of H3K27me3 scCUT&Tag profile sorting in Supplementary Figure 6c

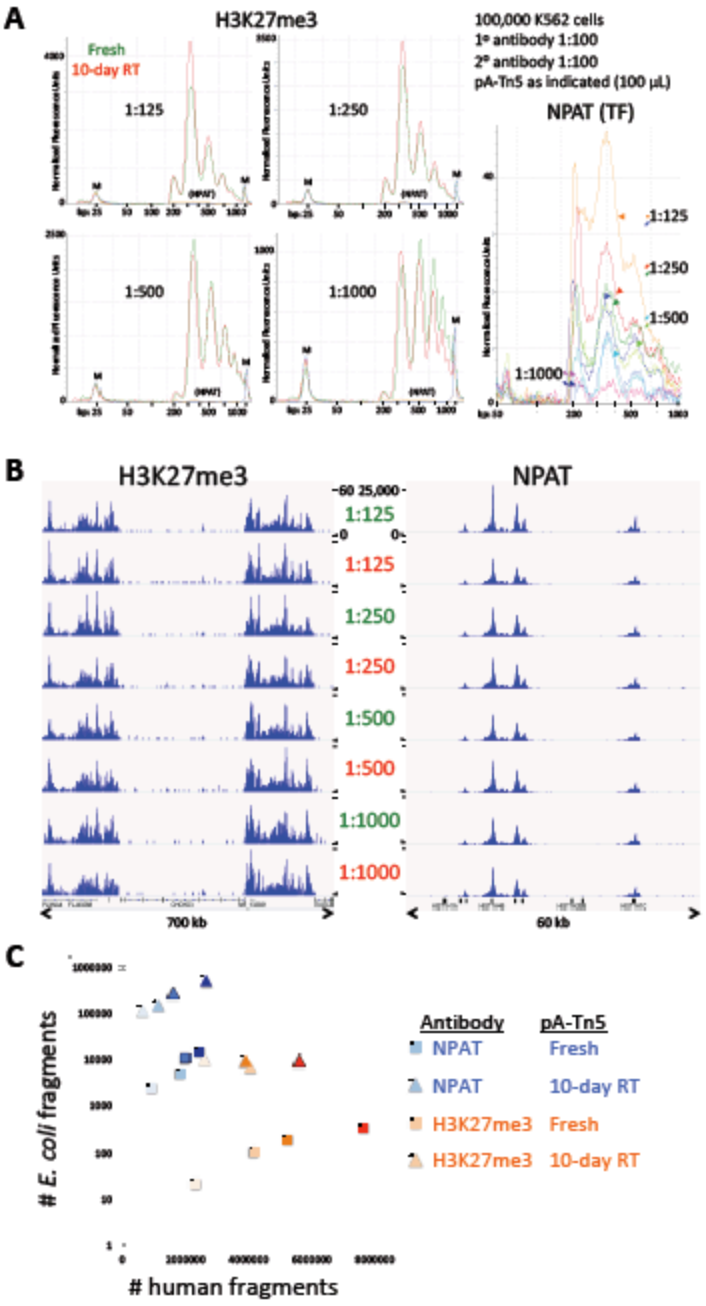


Figure 3.12: pA-Tn5 complex is stable at room temperature.
a) pA-Tn5 in glycerol storage buffer was loaded with adapters and stored at -20oC. A 5 µL aliquot was removed to a PCR tube and held for 10 days at room temperature (RT). CUT&Tag was performed by splitting the samples into 8 50 µL aliquots following the secondary antibody incubation, washing twice in Dig-wash buffer, then incubating 1 hr at RT with serial dilutions of either Fresh or RT pA-Tn5 in Dig-med buffer as indicated. After three Dig-med washes, cells were resuspended in Dig-med + 10 mM MgCl2 and incubated 1 hr 37oC for tagmentation.
DNA was extracted and 70% of the total DNA was subjected to 14 PCR cycles. After a single 1.1X Ampure-bead clean-up, DNA was eluted with 25 µL 10 mM Tris-HCl pH 8, and Tapestation D-1000 analysis was performed on a 2 µL sample. Markers (M): lower = 25 bp and upper = 1500 bp. Whereas

H3K27me3 is an abundant histone modification, NPAT is a transcription factor (TF) that is specific for the histone loci on human chromosomes 1 and 6, and therefore has very few target sites in the genome. b) Examples of tracks from serial dilution of pA-Tn5 (50 µL volume). All tracks for each antibody are at the same genome-normalized scale, 0-60 for H3K27me3 around the CHCHD3 locus (Chr7:132,280,000-132,980,000) and 0-25,000 for NPAT around part of a histone cluster (Chr6:26,010,000-26,070,000). c) E. coli carry-over DNA levels increase by 1-2 orders-of-magnitude during 10-day room temperature incubation of pA-Tn5. Counts of human and E. coli fragments are shown for the samples in panel B, where shading indicates serial dilutions from 1:125 (dark) to1:1000 (light).

Chapter 4

Single-cell CUT&Tag analysis of chromatin modifications in differentiation and tumor progression

Modified from an article originally published in *Nature Biotechnology:*
Steven J. Wu, Scott N. Furlan, Anca B. Mihalas, Hatice S. Kaya-Okur, Abdullah H. Feroze, Samuel N. Emerson, Ye Zheng, Kalee Carson, Patrick J. Cimino, C. Dirk Keene, Jay F. Sarthy, Raphael Gottardo, Kami Ahmad, Steven Henikoff & Anoop P. Patel. *Nature Biotechnology* volume 39, pages 819–824 (2021)

Abstract

Methods for quantifying gene expression (Tanay & Regev, 2017) and chromatin accessibility (Klemm et al., 2019) in single cells are well established, but single-cell analysis of chromatin regions with specific histone modifications has been technically challenging. In this study, we adapted the CUT&Tag method (Kaya-Okur, 2019) to scalable nanowell and droplet-based single-cell platforms to profile chromatin landscapes in single cells (scCUT&Tag) from complex tissues and during the differentiation of human embryonic stem cells. We focused on profiling polycomb group (PcG) silenced regions marked by histone H3 Lys27 trimethylation (H3K27me3) in single cells as an orthogonal approach to chromatin accessibility for identifying cell states. We show that scCUT&Tag profiling of H3K27me3 distinguishes cell types in human blood and allows the generation of cell-type-specific PcG landscapes from heterogeneous tissues. Furthermore, we used scCUT&Tag to profile H3K27me3 in a patient with a brain tumor before and after treatment, identifying cell types in the tumor microenvironment and heterogeneity in PcG activity in the primary sample and after treatment.

Introduction

Substantial portions of the genome are actively repressed to create barriers between cell type lineages during development (Lee, 2006). In particular, trimethylation on lysine 27 of histone H3 (H3K27me3) in nucleosomes by PcG proteins is crucial for gene silencing during normal differentiation and, thus, for maintaining cell identity (Laugesen & Helin, 2014). Conversely, derangements in PcG silencing permit aberrant gene expression and disease (Sparmann & Lohuizen, 2006). Therefore, methods for assaying silenced chromatin can provide insights into a variety of processes, ranging from normal development to tumorigenesis.

Scalable methods for assessing silenced chromatin at the single-cell level have not been widely available. We set out to use chromatin profiling of single cells to assess gene silencing and to develop a framework for analysis. Our approach builds on Cleavage Under Targets and Tagmentation (CUT&Tag), which uses specific antibodies to tether a Tn5 transposome at the sites of chromatin proteins in isolated

cells or nuclei. Activation of the transposome then tagments genomic loci with adapter sequences that are used for library construction and deep sequencing, thereby identifying binding sites for any protein where a specific antibody is available (Kaya-Okur, 2019). Our earlier work demonstrated that CUT&Tag profiling of the H3K4me2 histone modification efficiently detected gene activity, much like assay for transposase-accessible chromatin sequencing (ATAC-seq), whereas H3K27me3 profiling detected silenced chromatin that might be epigenetically inherited (Kaya-Okur, 2019).

Results

scCUT&Tag resolves distinct cell types and maps repressive chromatin domains in early hESC development

To determine whether single-cell chromatin landscapes were sufficient to distinguish different cell types, we performed CUT&Tag on H1 human embryonic stem cells (hESCs) using an anti-H3K27me3-specific antibody in bulk and then distributed single cells for polymerase chain reaction (PCR) and library enrichment on the nanowell-based ICELL8 system (Figure 4.1a). We compared this to previously published H3K27me3 scCUT&Tag profiles of K562 cells and hESCs3 to determine whether standard approaches to single-cell clustering could distinguish cell types based on H3K27me3 signal. As PcG domains typically span more than 10 kilobases (kb), we grouped read counts in 5-kb bins across the genome and used this for latent sematic indexing (LSI)-based dimensionality reduction and uniform manifold approximation and projection (UMAP) embedding, followed by standard Louvain clustering using the ArchR package (Granja, 2021) (Methods). After quality control filtering (Methods and Supplementary Figure 4.1a-g), UMAP embedding clearly separated 100% of 804 hESCs with a median of 375 unique fragments from 908 K562 cells with a median of 6,064 unique fragments independent of batch effects (Figure 4.1b). Interestingly, hESCs had 6% of the number of unique fragments when compared to K562 cells (Figure 4.2f). This demonstrates that stem cells have lower global H3K27me3 levels than more differentiated cell types (Hawkins et al., 2010). Despite downsampling the number of unique fragments per cell to the same median value for both datasets, H3K27me3 signal still readily distinguished the two cell types (Figure 4.2g), confirming that clustering was driven by differences in H3K27me3 signal and not the number of unique fragments.

Cellular determination and differentiation proceed by a controlled sequence of gene activation and gene repression. To study gene silencing during development, we differentiated hESCs toward definitive endoderm (Chu, 2016). We confirmed differentiation by immunofluorescence staining of stage-specific transcription factors (TFs) (Figure 4.3a). UMAP embedding of 1,830 scCUT&Tag H3K27me3 profiles with a median of 279 fragments revealed a developmental trajectory, independent of batch effect (Figure 4.3b), from hESC to definitive endoderm (Figure 4.1c) that was punctuated by stem-like states on days 1–2 followed by a rapid progression toward differentiation on days 3–5. To determine if changes in chromatin silencing correspond to changes in gene expression, we examined known markers of stem cells and endoderm differentiation in single-cell aggregate profiles from each day. Overall, H3K27me3

signal at a marker gene was inversely correlated with expression based on a published single-cell RNA sequencing (scRNA-seq) dataset (Chu, 2016). Stem cell markers, such as SOX2, KLF4 and FOXD3, were expressed in hESCs and lacked H3K27me3 but were silenced as differentiation proceeded (Figure 4.1d). Between days 2 and 3, hESCs transition into a mesendoderm state (characterized by expression of TBXT, MSX2 and PDGFRA) in which they have the developmental potential to become either mesoderm or endoderm (Chu, 2016). This is illustrated in our data between days 2 and 3 where chromatin silencing at mesoderm markers was lower (Figure 4.1d). As differentiation proceeded, endoderm markers, such as FOXA2, SOX17 and PRDM1, became active and lost H3K27me3 signal (Figure 4.1d). Finally, markers of ectoderm (PAX6 and LHX2) were not expressed and accumulated H3K27me3, consistent with silencing of these loci (Figure 4.3c). Pseudo-temporal ordering of single cells recapitulated our real-time results (Figure 4.3d).

scCUT&Tag for H3K27me3 readily identifies major subtypes in PBMCs

Having established that scCUT&Tag readily identifies dynamic changes in chromatin silencing, we next sought to determine whether chromatin profiles could distinguish cell types in a more complex tissue. To do so, we adapted scCUT&Tag to the 10x Genomics microfluidics platform and profiled H3K27me3 in mixed peripheral blood mononuclear cells (PBMCs) collected from two healthy donors. Briefly, we performed scCUT&Tag in bulk on 1 million cells and then loaded two lanes of a 10x Genomics microfluidic chip with 10,000 nuclei each to obtain technical replicates (Figure 4.4). We implemented a chromatin silencing score (CSS), which uses the gene activity score model in ArchR (Granja, 2021) to create a proxy for the overall signal associated with a given locus. Quality control filtering resulted in 9,917 cells with a median of 1,110 unique fragments per cell for which we performed dimensionality reduction and embedding as described above (Figure 4.5a). The median number of reads falls in the range expected for cell type variation, in spite of the platform differences in our study.

We then set out to identify the major cell types in the data using two methods. We first downsampled publicly available bulk H3K27me3 chromatin immunoprecipitation followed by sequencing (ChIP-seq) data (ENCODE) and used the UMAP transform function to ‘project’ the ChIP-seq data onto our UMAP embedding as previously described (Granja, 2019) (Figure 4.5a). We used the CSS score to identify cell-type-specific marker genes that showed a lack of H3K27me3 enrichment because active genes will have a low CSS. Therefore, we would expect a low CSS for a cell-type-specific marker gene in the cluster that corresponded to that cell type (Figure 4.5b). Overall, cluster identification by CSS annotation matched our assignments by ChIP-seq projection (Figure 4.5a) and distinguished major cell types in unsorted PBMCs, including those of lymphoid (T cell, natural killer cell and B cell) and myeloid (monocyte) lineages. We recovered the proportions of major cell types within the range of normal adult blood (Supplementary Table 1). Using this method, we can, therefore, generate cell-type-specific PcG landscapes across heterogeneous cell types within a sample, obviating the need for physical cell sorting and minimizing confounders such as batch effect, read depth or sample heterogeneity (Figure 4.4). This

allowed us to identify the top differentially PcG-silenced loci across the major cell types in PBMCs (Figure 4.5b). We also profiled PBMCs with the active mark H3K27ac and recovered the major cell types in a similar proportion as H3K27me3 scCUT&Tag (Figure 4.6 and Figure 4.7).

We next demultiplexed each biological donor using souporcell. In brief, the algorithm identifies genotypic differences between single cells by variant calling aligned reads (Heaton, 2020). The variant calls can also be used to identify multiplets. Using this method, we were able to differentiate cells from each donor (Figure 4.5c). Clustering was not driven by donor-specific effects but, rather, by cell type differences (Figure 4.4c).

scCUT&Tag data for H3K27me3 for a human glioblastoma primary and relapse sample

Having established that scCUT&Tag can profile developmental systems and heterogeneous tissues, we used scCUT&Tag to interrogate PcG-based clustering in glioblastoma, a human central nervous system tumor that is known to have a heterogeneous microenvironment (Bhaduri, 2020), exhibit intratumoral heterogeneity (Patel, 2014) and have pseudo-hierarchical organization that mimics development (Bhaduri, 2020; Couturier, 2020; Wang, 2017). In this tumor type, changes in PcG chromatin silencing can mediate emergence of resistant cell populations (Liau, 2016).

We profiled H3K27me3 in 1,311 single cells (3,643 median fragments per cell) using the 10x scCUT&Tag workflow from a primary glioblastoma that had been snap-frozen shortly after surgical removal. We distinguished four major cell populations within the sample (Figure 4.8a). To annotate clusters, we constructed CSS of previously defined marker loci (Bhaduri, 2020) and annotated clusters that correspond to microglia (Cluster 1, low CSS at the PTPRC gene), neurons (Cluster 3, low CSS at RBFOX3), oligodendrocytes (Cluster 4, low CSS at MOBP) and other neural lineage cells, including tumor cells (Cluster 4, low CSS at SOX2) (Figure 4.8b). To confirm cluster annotations, we projected CUT&RUN bulk data from a glioma stem cell line (UW7gsc) derived from the same patient, two established neural stem cell lines (U5 and CB660)17, and ENCODE (ENCODE Project Consortium, 2012) ChIP-seq bulk data for monocytes (proxy for microglia) and astrocytes. Projection onto the scCUT&Tag tumor sample embedding confirmed CSS annotations (Figure 4.8c). UW7gsc projected to the center of the largest cluster, presumably made up of tumor cells. The astrocyte data projected to a smaller satellite cluster within the neural lineage cells. The neural stem cell line data localized to both the tumor cell cluster and the astrocyte cluster (Figure 4.9). This might reflect spontaneous differentiation of neural stem cells toward the astrocyte lineage in vitro or reflect subtle changes in cell state, such as lineage priming (Llorens-Bobadilla, 2015).

To understand how the tumor changed with treatment, we performed scCUT&Tag profiling for H3K27me3 for a relapse sample obtained via rapid autopsy from the patient 5 months after surgery and radiation therapy. Application of quality control metrics followed by low-dimensional embedding identified four distinct cell types in the relapse sample (Figure 4.10a). Projection of the 1,168 autopsy single-cell

profiles (16,232 median fragments per cell) onto the primary tumor UMAP embedding allowed cell type identification, including 71 cells that co-localized to the tumor cell cluster (Figure 4.10b).

The low number of cells in the autopsy tumor specimen that met quality control standards limited the biological conclusions that we could draw from these data independently. As such, we chose to consider the 71 autopsy cells in the context of the 640 primary tumor cells by co-embedding them together and analyzing their relationship to each other. After batch correction, we identified four clusters within the tumor cell data with distinct H3K27me3 profiles (Figure 4.8e, left). Examining the distribution of cell states across the two time points, we noted an enrichment for Cluster T1 in the relapse specimen (Figure 4.8d right). The relapse tumor cells had higher background signal when compared to the primary tumor cells as determined by FRiP analysis (Figure 4.10c). To further confirm that the relapse cells were most similar to Cluster T1, we characterized reads in relapse cells that were present in genomic regions that most significantly distinguished the primary tumor clusters (Figure 4.10d). This analysis confirmed similarity of the relapse tumor cells to Cluster T1. Gene set enrichment analysis using the CSS matrix identified potential programs silenced (positive enrichment scores) and derepressed (negative enrichment scores) in this cluster. Interestingly, the Verhaak_glioblastoma_proneural gene set appears to be silenced in the resistant cell cluster (Figure 4.11), consistent with the idea that tumor evolution might induce a proneural-to-mesenchymal shift (Segerman, 2016). In contrast, low CSS was observed at gene sets with high CpG content that are marked by H3K27me3 in whole brain (Meissner, 2008). The lack of H3K27me3 signal in this tumor cluster suggests that the PcG landscape of glioblastoma cells resembles a stem-like state rather than a terminally differentiated state (Rheinbay, 2013).

We next wanted to understand the relationship between the cell clusters. Clusters T1, T2 and T4 exist along a continuum, whereas Cluster T3 is separated from the main tumor cell group. We focused on whether TF programs are differentially silenced across Clusters T1, T2 and T4. H3K27me3 domains are broad, spanning 10–100 kb and covering many genes, enhancers, promoters and intervening regions. Therefore, to limit motif searching to potential regulatory elements within H3K27me3 domains, we used single-cell ATAC-seq data (Figure 4.12) to annotate enhancers and promoters in tumor cell subclusters based on accessible chromatin. We then calculated TF motif enrichments and depletions in this set of curated genomic regions based on H3K27me3 signal. We examined motif deviations ordered over two pseudotime trajectories that started with Cluster T1 (presumed stem-like cluster) and ended in either Cluster T4 (Trajectory 1) or Cluster T2 (Trajectory 2) (Figure 4.8e, left). Motif deviations (n = 132) were ordered according to pseudotime, identifying silenced motifs that spanned Cluster T1 to Cluster T2 and Cluster T4 (Figure 4.8e, right). At the apex of the trajectories, motif silencing was shared and included motifs for TFs such as NEUROD1, SNAI2 and TCF12 (Figure 4.8f, left column). At intermediate pseudotime points, there were silenced motifs specific to Trajectory 1 (NR1DA2) or Trajectory 2 (ETV5) or shared by both (RFX4) (Figure 4.8f, middle column). As pseudotime proceeded, Trajectory 1 showed evidence of HES5 motif silencing, whereas Trajectory 2 showed GATA6 motif silencing. Interestingly, the

DNMT1 motif was strongly silenced across both pseudotime endpoints, concordant with the idea that PcG silencing of DNMT1-enriched promoters and enhancers is a common feature of differentiation (O'Neill, 2018) (Figure 4.8f, right column).

Discussion

Fundamentally, we showed here that repressive chromatin can be used to identify cell states a priori from heterogeneous normal and diseased tissues. This approach has far-reaching applications, including generation of cell-type-specific chromatin atlases from archival tissue in a manner that does not require sorting of pure populations. We focused primarily on a single chromatin mark, but this method can, in theory, be applied to any histone modification or DNA-binding protein for which an antibody is available. As such, developing complete chromatin landscapes of complex tissues and disease states using scCUT&Tag will help decode the complex epigenetic machinery underlying gene expression. Broadly, our method for performing histone mark-specific single-cell analysis adds to the growing list of single-cell 'omic' methods that can be used to understand heterogeneous cell populations.

Methods

Biological material

H1 hESCs were purchased from WiCell (cat. no. WA01, lot no. WB35186). We used the following antibodies: guinea pig anti-rabbit IgG (heavy and light chain) antibody (antibodies-online, ABIN101961), H3K27me3 (Cell Signaling Technology, cat. no. 9733), H3K27ac (Millipore Sigma, cat. no. MABE647, lot no. VP1901251), SOX17 (R&D Systems, AF1924, lot KGA0916121), OCT4 (Abcam, ab109183, lot gr120970-6) and H3K4me2 (Upstate Biotechnology, 07–030, lot 26335) The fusion enzyme pA-Tn5 was generated as previously described3.

hESC culture conditions

H1 hESCs were maintained on Corning Matrigel hESC-Qualified Matrix (Corning, no. 354277) at 37 °C in mTeSR1 from STEMCELL Technologies (cat. no. 85850) with daily medium replacement. When cell aggregates were 80% confluent, they were released using ReLeSR (STEMCELL Technologies, no. 05872) per manufacturer instructions and incubated at 37 °C for 3–5 min. Cells were released into a small volume of complete medium by tapping of growth plate, and aggregates reduced in size by gentle pipetting and passaged to desired ratio.

hESC differentiation protocol

hESCs were differentiated to definitive endoderm using the STEMdiff Definitive Endoderm Kit (cat. no. 05110). The full protocol is available from STEMCELL Technologies (https://cdn.stemcell.com/media/files/pis/29550-PIS_2_1_0.pdf?_ga=2.73376023.564267965.1597964514-138601152.1597964514). Briefly, hESCs at

80% confluence were harvested using Gentle Cell Dissociation Reagent (STEMCELL Technologies, cat. no. 07174) and reseeded in a single-cell manner on Matrigel plates. This was done daily for 5 d. Every 24 h after a new differentiation, culture was started, and cells were incubated with DE differentiation medium according to the manufacturer's guidelines. On the fifth day, all five time points were harvested simultaneously using Accutase (STEMCELL Technologies, cat. no. AT104-500). Immunofluorescence was used to confirm differentiation as previously described (Meers, Janssens, et al., 2019) with primary antibodies SOX17 1:250 dilution (R&D Systems, AF1924, lot KGA0916121) and OCT4 1:100 dilution (Abcam, ab109183, lot gr120970-6) and secondary antibodies donkey anti-goat Rhodamine Red 1:1,000 dilution (Jackson ImmunoResearch, cat. no. 705-295-147) and goat anti-rabbit-Cy5 1:1,000 dilution (Jackson ImmunoResearch, cat. no. 111-175-144).

PBMC acquisition and processing

Healthy adult donors at the University of Washington underwent venipuncture, and blood was collected using heparin-containing vacutainer tubes after consenting to participate in our study (Institutional Review Board (IRB) protocol no. STUDY00008678). Additional PBMC specimens were obtained from consented donors at the Fred Hutchinson Cancer Research Center (IRB no. 0999.209). Mononuclear cells were harvested from peripheral blood using gradient centrifugation. Cells were then washed twice with PBS and captured as outlined below.

Brain tumor specimen acquisition, processing and culture

Adult patients at the University of Washington provided pre-operative informed consent to take part in the study in all cases after approved IRB protocol (no. STUDY00002162). Fresh tumors were collected directly from the operating room at the time of surgery and either taken fresh or snap-frozen immediately after removal in liquid nitrogen. Histopathologic diagnosis was confirmed by a board-certified neuropathologist. Fresh tissue was enzymatically dissociated using a papain-based brain tumor dissociation kit (Miltenyi Biotec) as per the manufacturer's protocol. Cells were then cultured on laminin-coated plates in DMEM/F12 supplemented with 1× N2/B27 and 1% penicillin–streptomycin. Cultures were passaged as needed when confluent and considered stable after three serial passages. Cell line UW7gsc was used for this study at passage 3. Autopsy tissue was collected with a post-mortem interval of approximately 8.75 h after informed consent with a waiver from the University of Washington IRB. Tissue was snap-frozen in liquid nitrogen-cooled isopentane. Tumor regions were sampled based on gross examination of brain sections and processed as outlined below.

Nuclei preparation from brain tumor specimens

Frozen tissue was processed to nuclei using the 'Frankenstein' protocol from protocols.io. Briefly, snap-frozen glioblastoma tissue was thawed on ice and minced sharply into <1-mm pieces. Next, 500 µl

of chilled Nuclei EZ Lysis Buffer (Millipore Sigma, NUC-101, no. N3408) was added, and tissue was homogenized 10–20 times in a Dounce homogenizer. The homogenate was transferred to a 1.5-ml Eppendorf tube, and 1 ml of chilled Nuclei EZ Lysis Buffer was added. The homogenate was mixed gently with a wide-bore pipette and incubated for 5 min on ice. The homogenate was then filtered through a 70-μm mesh strainer and centrifuged at 500g for 5 min at 4 °C. Supernatant was removed, and nuclei were resuspended in 1.5 ml of Nuclei EZ Lysis Buffer and incubated for 5 min on ice. Nuclei were centrifuged at 500g for 5 min at 4 °C. After carefully removing the supernatant (pellet might be loose), nuclei were washed in wash buffer (1× PBS, 1.0% BSA and 0.2 U μl−1 of RNase Inhibitor). Nuclei were then centrifuged and resuspended in 1.4 ml of wash buffer for two additional washes. Nuclei were then filtered through a 40-μm mesh strainer. Intact nuclei were counted after counterstaining with Trypan blue in a standard cell counter.

Chromatin profiling: scCUT&Tag using the ICELL8 system/protocol

scCUT&Tag for the ICELL8 was carried out as previously described3. In brief, approximately 250,000 hESCs (for each time point) were processed by centrifugation between buffer exchanges at 600g for 3 min and in low-retention tubes. Cells were collected and washed with 1 ml of wash buffer (20 mM HEPES pH 7.5, 150 mM NaCl, 0.5 mM spermidine and 1× protease inhibitor cocktail) at room temperature. Cells were incubated in antibody diluted 1:50 in NP40-Digitonin Wash Buffer (0.01% NP40 and 0.01% digitonin in wash buffer) overnight. This wash buffer permeabilized the cells and released nuclei. Permeabilized nuclei were then rinsed once with NP40-Digitonin Wash buffer and incubated with anti-rabbit IgG antibody (1:50 dilution) in 1 ml of NP40-Digitonin Wash buffer on a rotator at room temperature for 30 min. Nuclei were washed twice with NP40-Digitonin Wash buffer and incubated with 1:100 dilution of pA-Tn5 in NP40-Dig-med-buffer (0.01% NP40, 0.01% digitonin, 20 mM HEPES pH 7.5, 300 mM NaCl, 0.5 mM spermidine and 1× protease inhibitor cocktail) for 1 h at room temperature on a rotator. Cells were washed two times with NP40-Dig-med-buffer and resuspended in 150 μl of

tagmentation buffer (10 mM MgCl2 in NP40-Dig-med-buffer) and incubated at 37 °C for 1 h. Tagmentation was stopped by adding 50 µl of 4× Stop Buffer (40.4 mM EDTA and 2 mg ml−1 of DAPI), and the sample was held on ice for 30 min. Samples were then strained through a 10-µm cell strainer to remove clumps of cells.

The SMARTer ICELL8 single-cell system (Takara Bio, cat. no. 640000) was used to array single cells as previously described3. Briefly, cells were loaded onto a source plate and dispensed into a SMARTer ICELL8 350v Chip (Takara Bio, cat. no. 640019) at 35 nl per well. The chip was then spun down at 300g for 5 min. Imaging on a DAPI channel confirmed the presence of single cells in specific wells. Non-single cell wells were excluded from downstream reagent dispenses. To index the whole chip, 72 × 72 i5/i7 unique indices (5,184 microwells total) were dispensed at 35 nl in wells that contained single cells, followed by two dispenses of 50 nl (100 nl total) of 2× NEBNext High-Fidelity 2× PCR Master Mix (NEB, M0541L). The chip was sealed and spun down at 2,250g for 3 min after each dispense. The PCR on the chip was performed with the following protocol: 5 min at 72 °C and 2 min at 98 °C, followed by 15 cycles of 10 s at 98 °C, 30 s at 60 °C and 5 s at 72 °C, with a final extension at 72 °C for 1 min.

Quality control (ICELL8)

The ICELL8 has a built-in imaging system that filters out wells that do not contain a single cell. Thus, empty wells without cells, with more than one single cell and with doublets are removed. Subsequently, we filtered single cells with fewer than 100 unique fragments to remove spurious barcodes that can be attributed to an overflow of dispensed PCR material.

A drawback of leveraging a hyperactive transposon in a fusion enzyme to target specific chromatin compartments is that the Tn5 has a high binding affinity for accessible chromatin, the basis of ATAC-seq. Previously, it was shown that this artifact is highly dependent on the concentration of salt in subsequent washes after fusion enzyme binding3. To identify whether our single-cell samples exhibited this artifact, we mapped the percent of reads in each single cell that fell into H3K27me3, H3K4me2 or ATAC-specific peaks (Figure 4.2c). The degree in which repressive H3K27me3 marked chromatin and active accessible chromatin ATAC-seq signal overlapped was minimal, as expected, whereas an active mark, H3K4me2, had a higher degree of overlap with ATAC-seq data. Correlations of aggregate versus bulk profiles across the 5-kb genome tiles showed similar results (Figure 4.2b).

As an initial test, we wanted to evaluate the robustness of scCUT&Tag by comparing it to single-cell ATAC-seq. Therefore, we chose the histone modification K4me2 that was shown to provide similar output to ATAC-seq. A representative genomic track comparing bulk, aggregate and single-cell profiles for K4me2 in H1 and K562 cells revealed the high-quality resulting data (Figure 4.2a). A low-dimensional embedding, UMAP, clearly separates K562 cells (n = 807) from hESCs (n = 317) (Figure 4.2d).

Projections of published single-cell ATAC-seq data (GSE99172) onto our scCUT&Tag embedding align with cell-type-specific clusters (Figure 4.2e).

Chromatin profiling: scCUT&Tag using the 10x Genomics system

CUT&Tag was performed with an anti-H3K27me3 antibody (Cell Signaling Technology, no. 9733, dilution 1:100) or anti-H3K27ac (MABE647, dilution 1:100) with 1 million cells, as previously published3. Guinea pig anti-rabbit IgG secondary antibody (ABIN1011961) was used at 1:100 dilution. Adaptation to the 10x workflow was performed as follows. For all samples except the PBMC mixing experiment, the nuclei were spun down at 600g for 3 min after the pA-Tn5 binding step. After counting, they were resuspended in 1× Diluted Nuclei Buffer at 2,500 nuclei per microliter. The nuclei were then prepared for transposition per the 10x Genomics single-cell ATAC-seq protocol (SingleCell_ATAC_ReagentKits_v1.1_UserGuide_RevD). All steps beginning with 1.1, 'Prepare Transposition Mix', were performed according to 10x Genomics standard protocol. Libraries were sequenced using an Illumina NovaSeq 6000.

For the PBMC mixing experiment, the nuclei were tagmented in high salt (300 mM) as per published protocol3. After tagmentation, BSA was added to a final concentration of 1% and nuclei were centrifuged at 600g for 3 min and then resuspended in 1× Diluted Nuclei Buffer (10x Genomics, PN-2000207) at 2,500 nuclei per microliter. The 10x Genomics single-cell ATAC-seq protocol (SingleCell_ATAC_ReagentKits_v1.1_UserGuide_RevD) was used with the following modifications. For Step 1.1, 'Prepare Transposition Mix', 7 µl of ATAC buffer, 3 µl of low TE buffer (10 mM Tris pH 8.0, 0.1 mM EDTA) and 5 µl of stock nuclei solution were mixed together, omitting the ATAC enzyme, as tagmentation had already been performed. All remaining steps, beginning with Step 2.0, 'GEM Generation and Barcoding', were performed according to 10x Genomics standard protocol. Libraries were sequenced using an Illumina NovaSeq 6000.

Data processing

Illumina.bcl files were demultiplexed and converted to FASTQ format using the cellranger-mkfastq function. Resulting FASTQ files were aligned to the hg38 genome, filtered for duplicates and counted using cellranger-atac. An output BED file of filtered fragment data containing the cell barcode was then read into ArchR (Granja, 2019) as fragment counts in 5-kb genome windows, which was used in all dimensionality reduction steps across all experiments. We used the ArchR (Granja, 2021) gene activity score to calculate our CSS as described above. We used LSI dimensionality reduction (Granja, 2021) using a TFIDF normalization function (Stuart, 2019), UMAP (Becht et al., 2018) low-dimensional embedding and clustering using a nearest neighbor graph25 performed on data in LSI space.

As the cell line/differentiation experiments used the ICELL8 platform, we did not remove multiplets, as this platform uses microscopic imaging to ensure single-cell capture. For droplet partitioning data, we used the following methods to ensure data quality. 1) We first visualized fragment length

distribution across clusters. We identified three clusters with nucleosomal banding distribution that was consistent with untethered transposition events (Figure 4.4b). 2) We then removed two clusters with high mean fragment counts. 3) We iteratively removed clusters that exhibited non-specific CSS. We accomplished this by calculating CSS significance across clusters using ArchR (Granja, 2021). Any cluster that did not have any genes that were significantly over-represented or under-represented using significance thresholds of false discovery rate < 0.01 and absolute fold change > 3 was removed. Bulk projection of downsampled ChIP-seq data was performed as follows. Raw sequence data aligned to hg38 (BAM files) were downloaded from ENCODE. Data were processed using ChomVAR27 by counting reads in 5-kb tiled genomes and subsequently used in the bulk projection function in ArchR. Shared nearest neighbor clustering was performed using Seurat from within ArchR. Single-cell projection was performed using a modified ArchR projection function that did not perform any manipulation of the input data before projection. Marker regions/genes for each group were calculated using the getMarkerFeatures function in ArchR. Pre-ranked GSEA (fgsea (Sergushichev et al., 2016)) was performed using the entire list of marker genes ranked by −log10(P value)/sign(fold change) with the complete MSigDB (Liberzon, 2011) set of gene lists. Peak set from single-cell ATAC-seq data (see below) was used as a custom annotation set, and motif deviations were calculated using the addDeviationsMatrix function in ArchR. Pseudotime trajectory was assigned with T1 as a root and Clusters T2 and T4 as an endpoint.

To perform variant calling, we first merged BAM output from cellranger-atac using a custom script (https://github.com/scfurl/mergeBams). We then used souporcell (Heaton, 2020) on the merged BAM, invoking the 'no_umi' and 'skip_remap' options. Sparse mixture model output from souporcell was log-normalized and colored by the genotype assignment.

Quality control and data processing for brain tumor ATAC-seq

Nuclei preparation from snap-frozen brain tumor tissue was performed as described above, and standard single-cell ATAC-seq workflow was performed as per manufacturer guidelines (10x Genomics). Sequencing data were processed using the cell ranger-atac package. An output BED file of filtered fragment data containing the cell barcode was then read into ArchR using 500-bp genome windows. We used LSI dimensionality reduction using a TFIDF normalization function, UMAP low-dimensional embedding and clustering using a nearest neighbor graph25 performed on data in LSI space. Tumor cells were identified as the largest cluster containing high gene activity scores for marker genes SOX2 and PTPRZ1. This cluster was used for peak calling using the MACS2 wrapper in ArchR with standard parameters.

External data

Data from the following identifiers were downloaded from the ENCODE portal (https://www.encodeproject.org) and Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/). For

Figure 4.2b and Figure 4.2a,f,g: GSE124557. For Figure 4.1d and Figure 4.3c: GSE75748. In addition, for the purposes of this study, hESC-differentiated time point 1.5 (scRNA-seq) was approximated to be day 2 in the GSE75748 dataset. For Supplementary Figure 4.3b,c,e: GSE99172, GSE99173, GSE124557 and GSE85330. For Figure 4.5b: ENCSR000ASK, ENCSR043SBG, ENCSR103GGR, ENCSR404MOX and ENCSR939JZW. For Figure 4.8c, the following data sets were used: ENCFF363TCY and ENCFF911MNN.

Data availability

Sequencing data are deposited in the Gene Expression Omnibus with accession code GSE157910. There are no restrictions on data use.

Author Contribution

S.J.W., S.N.F., A.B.M., H.K-O., A.H.F., S.N.E., and J.F.S. processed samples and performed experiments. S.J.W., S.N.F., Y.Z., R.G. and A.P.P. performed and/or provided input on data processing and analysis. K.C., P.J.C. and C.D.K. provided access to tissue samples and assisted with processing. S.J.W., S.N.F., K.A., S.H. and A.P.P. wrote the manuscript with input from all authors.

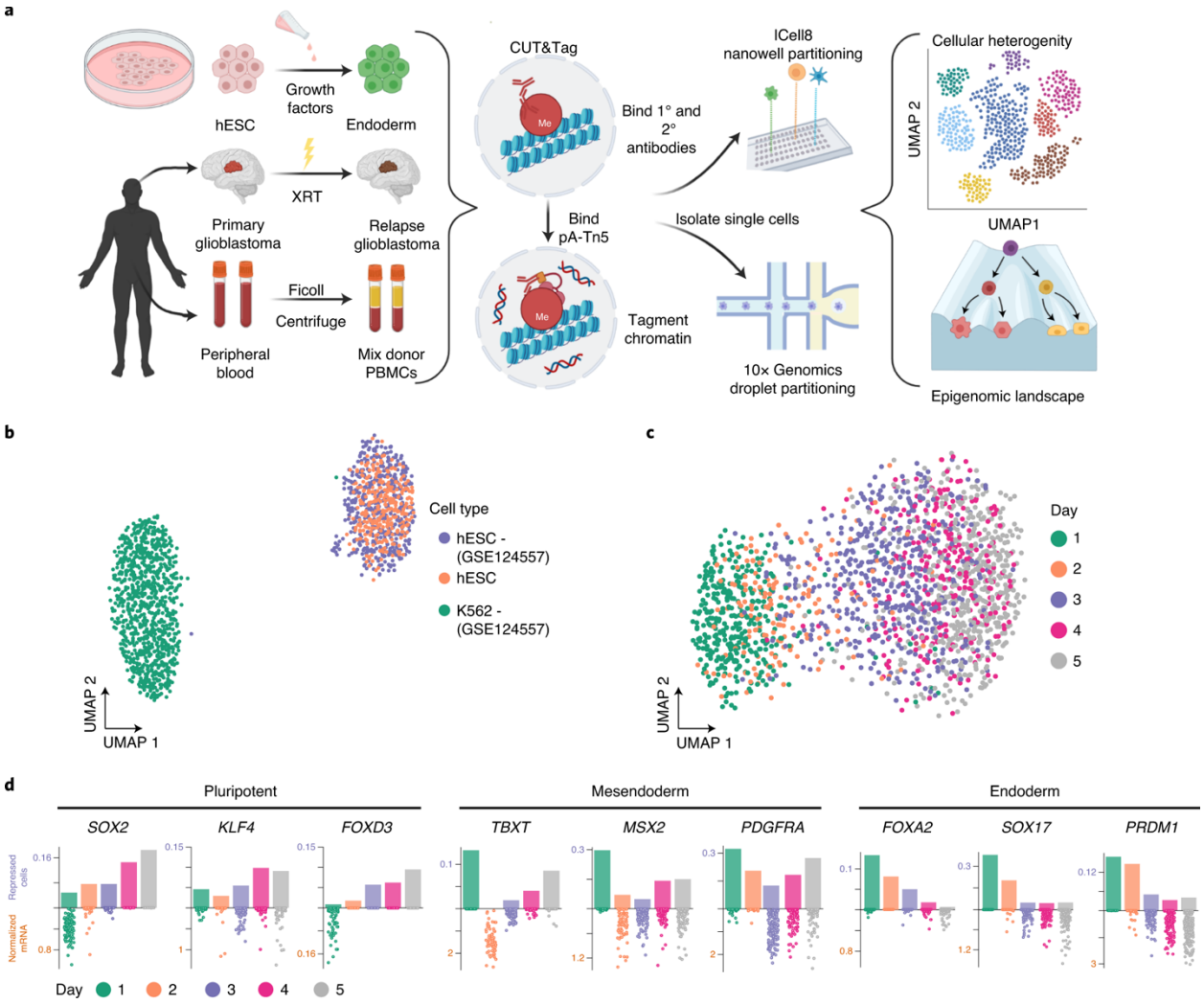


Figure 4.1: scCUT&Tag resolves distinct cell types and maps repressive chromatin domains in early hESC development.

a) Schematic of scCUT&Tag applied to nuclei isolated from cell culture, a model endoderm differentiaton system, blood cells and a human brain tumor. Single cells are then partitioned using either the 10x Genomics or ICELL8 microfluidic systems. b) UMAP embedding of scCUT&Tag for a repressive histone modification, H3K27me3, in K562 (n = 908) and hESC (n = 804) single cells. c) UMAP embedding of scCUT&Tag for a repressive histone modification, H3K27me3, in a 5-d differentiation time course from hESC to definitive endoderm (total n = 1,830). Cell types are colored according to the day along the time course in which they were harvested. d) Top, bar plot representing the percent of single cells (n = 350, 171, 474, 274 and 561 from days 1–5, respectively) that are repressed at each specific gene, where the upper axis corresponds to scCUT&Tag (percent of single cells repressed). Bottom, jitter plot depicting scRNA-seq for similar time points (n = 92, 66, 172, 138,and 188), where the lower axis corresponds to scRNA-seq (normalized messenger RNA counts from GSE75748). From left to right, well-known TF markers for pluripotent, mesendoderm and definitive endoderm cells. mRNA, messenger RNA.

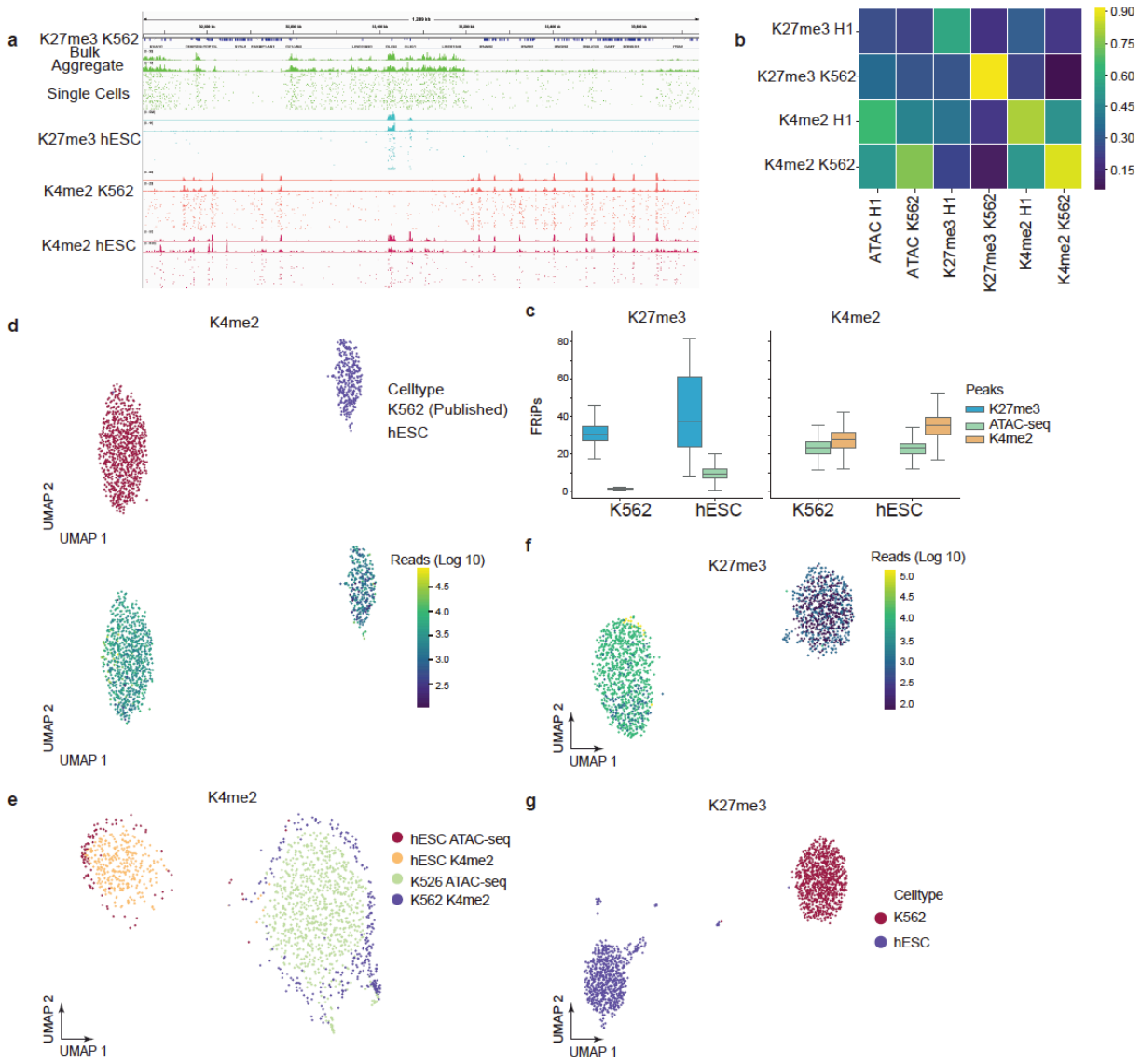


Figure 4.2: Quality control for scCUT&Tag in cell lines.
a) Representative chromatin landscapes across 1.2 Mb segment of the human genome for bulk, aggregate, and the top 50 single cells. b) A heatmap comparing the Pearson correlation of fragment counts in 5 kb bins across the genome for bulk and aggregate datasets. c) Boxplots illustrating the fraction of reads in peaks (FRiPs in percentages) (median=30.2, 1.2, 37.5, 9,4, 23.4, 27.7, 23.1, and 35, respectively) for each single cell (n=908, 804, 807, and 318 cells for each experiment and cell type, respectively) from peaks called in bulk datasets. The middle line is the median, the lower and upper lines correspond to the first and third quartile, and the upper and lower whiskers extend to the largest value no further than 1.5 × IQR (inter-quartile range). ATAC-seq datasets are used as a control for background. d) UMAP embedding of single cells for H3K4me2 shaded by their respective celltype (top) and unique number of reads (bottom). e) UMAP embedding of singles cells for H3K4me2. Published single cell ATAC-seq data from the same cell type are taken and projected onto the H3K4me2 low dimensional space. f) UMAP projection of H3K27me3 single cells shaded with the number of unique fragments per cell. g) UMAP projection of K27me3 single cells for K562 and hESC when K562 are downsampled to 365 reads per cell.

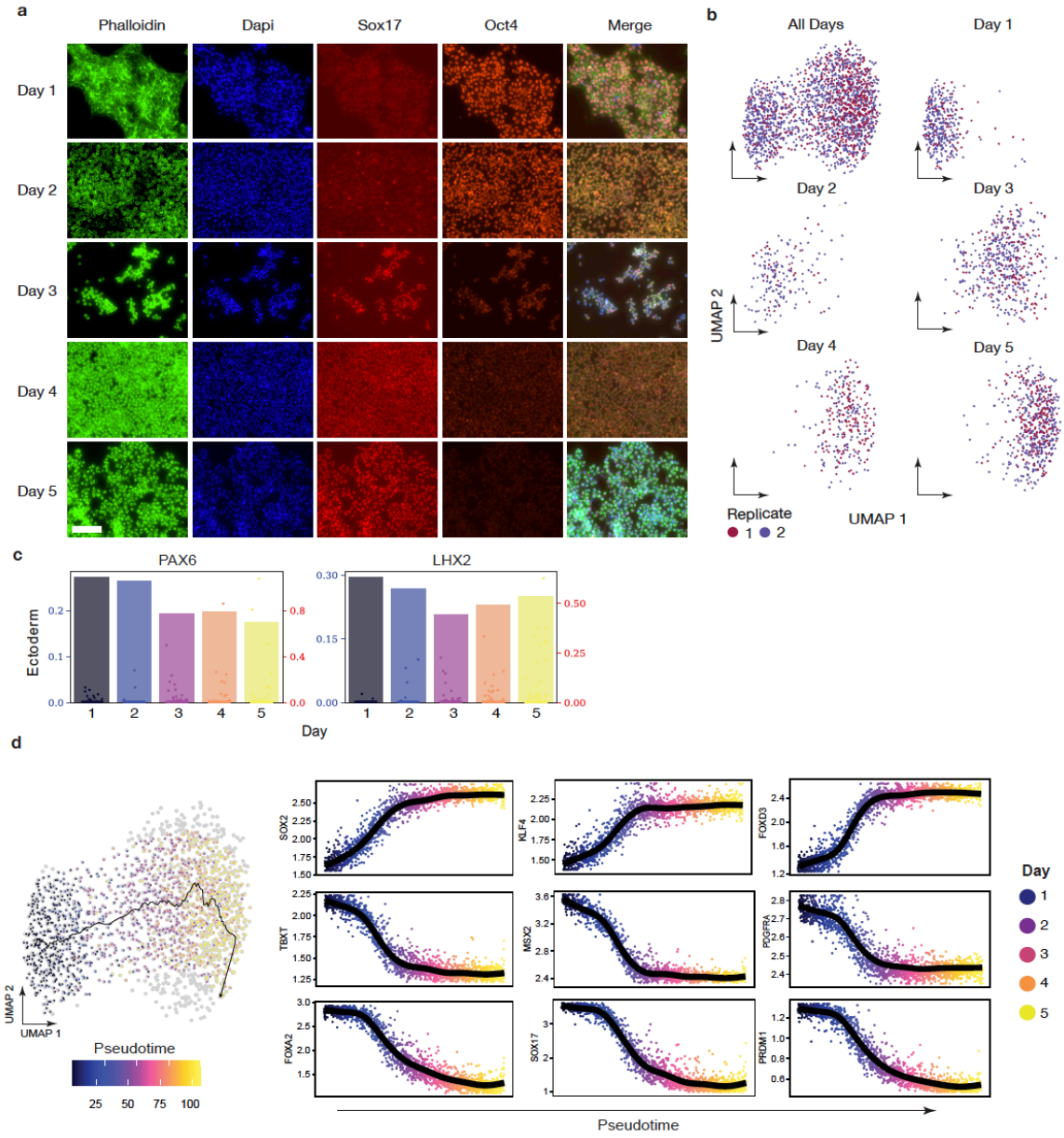


Figure 4.3: Quality control for scCUT&Tag in a five day time coarse of differentiated hESC.
a) A representative example of an immunofluorescence staining (n=3) for well-known TF markers that indicate either pluripotency (Oct4) or definitive endoderm (Sox17). Scale bar: 275 μm b) A grid plot illustrating UMAP projections for two biological replicates along different time points. c) A bar plot representing the percent of single cells that are repressed at each specific gene (n=350, 171, 474, 274, and 561 cells from day 1 to 5, respectively). The superimposed jitters depict scRNA-seq for a similar timepoint (n=92, 66, 172, 138, and 188 cells from day 1 to 5, respectively). The left axis corresponds to scCUT&Tag (percent of single cells repressed) and right corresponds to scRNA-seq (normalized mRNA counts). d) On the left a UMAP embedding showing the pseudotemporal ordering of single cells from a hESC to endoderm time course. On the right, dot plots of pseudo-time versus imputed chromatin silencing score at the gene of interest.

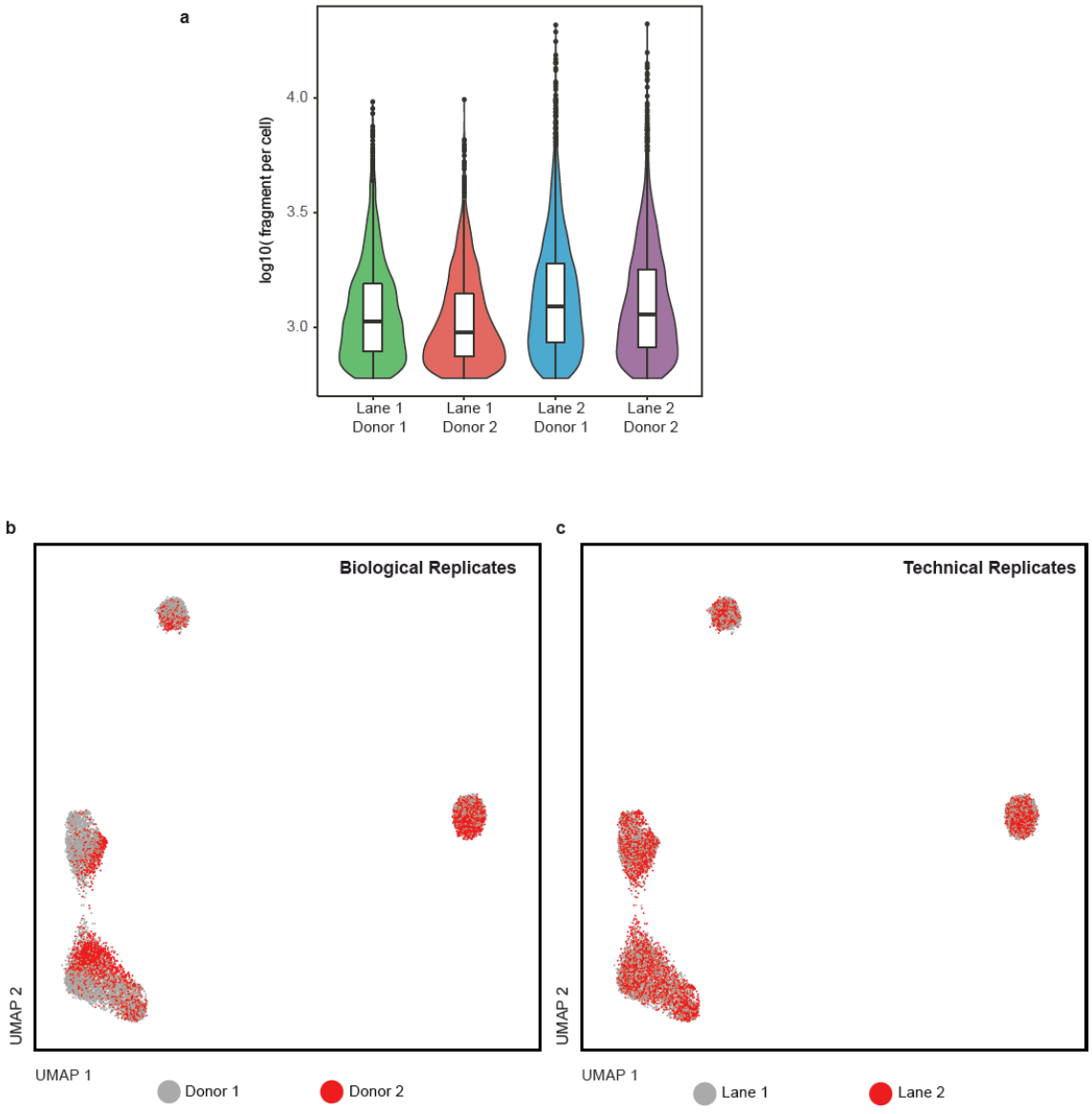


Figure 4.4: scCUT&Tag biological replicates (across donors) and technical replicates (across lanes). a) Violin plots depicting number of unique fragments (left to right, median=3.03, 2.98, 3.09, and 3.06) in each cell (left to right, n=2002, 1884, 3138, and 2892) within different biologic and technical replicates. Number of fragments recovered is similar across either different donors or across different 10X lanes. b) UMAP embedding colored by biologic replicates. Donors 1 and 2 overlap demonstrating minimal donor specific effects on clustering. c) UMAP embedding colored by technical replicates. Pooled PBMCs were run on two separate lanes of a 10X microfluidic chip. The two technical replicates overlap without replicate specific effects on clustering.

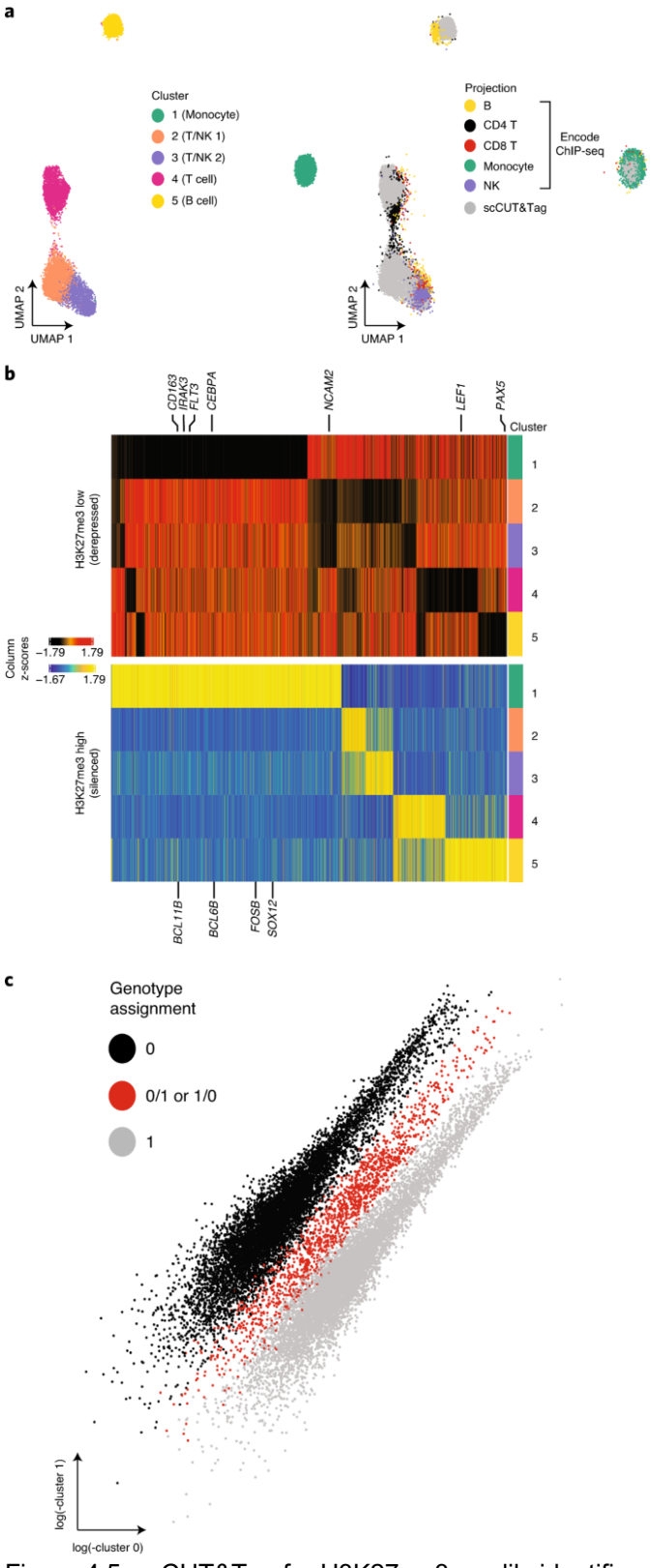


Figure 4.5: scCUT&Tag for H3K27me3 readily identifies major subtypes in PBMCs.

a) Left, UMAP embedding of single-cell data from PBMCs. Unsupervised clustering revealed five clusters. Right, UMAP projection of downsampled ChIPseq bulk data from primary sorted bulk datasets for major PBMC cell types (see Supplementary Methods for GSE citations) on single-cell CUT&Tag data on left. b) Heat map of genes with significantly low (top) or high (bottom) H3K27me3 signal (CSS) in each cluster (row). Fold change < -2 (top) or > 2 (bottom); $q < 0.05$ (both). Cell-type-specific genes are highlighted. c) Sparse mixture model clustering (using souporcell11) of genotype variant calls from the PBMC data colored by genotype assignment (before multiplet removal). NK, natural killer.

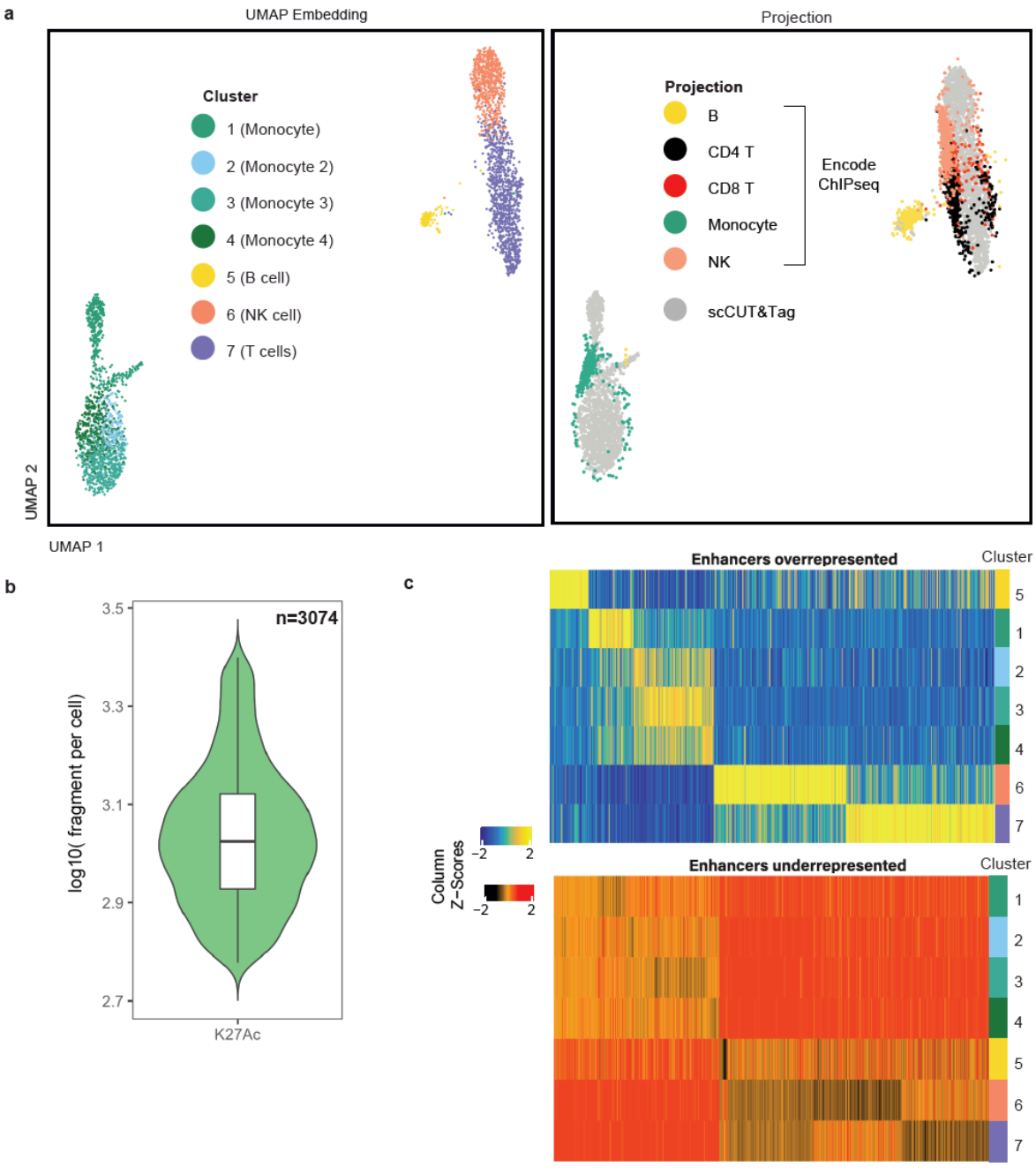


Figure 4.6: scCUT&Tag for H3K27ac readily identify major subtypes in PBMC.
a) Left - UMAP embedding of single cell data from PBMC. Unsupervised clustering revealed 7 clusters. Right - UMAP projection of downsampled ChIPseq bulk data from primary sorted bulk datasets for major PBMC cell-types on single cell CUT&Tag data on left. ENCODE accessions: ENCSR000ASJ - Monocytes, ENCSR000AUP - B cells, ENCSR007HLH - CD8 T, ENCSR138DOM - CD4, ENCSR391EQV - NK. b) Violin plot of fragments (median=3.02) per cell (n=3074). c) Heatmap of genes with significantly high (top) or low (bottom) H3K27ac signal in each cluster (row). Fold change > 2 (top) or < -2 (bottom); q-value < 0.05 (both).

a

PBMCs

Cluster (Cell Type)	H3K27me3	H3K27ac	Normal adult ranges
Mono	20.90%	47.60%	10-30%
NK or T	66.40%	48.30%	45-70% T 5-10% NK
B	12.70%	4%	5-15%

b

Glioblastoma

	scATAC-seq	scCUT&Tag
Microglia	20.70%	21.50%
Astrocytes	2.70%	3.50%
Tumor	50.40%	57.50%
Oligodendrocytes	4.40%	5.40%
Neurons	19.00%	12.10%

Figure 4.7: Comparison of major cell types in PBMC recovered.
a) A table comparing the proportions of major cell types in PBMCs identified by scCUT&Tag for H3K27me3 and H3K27ac to scATAC- seq and FACS. b) A table comparing the proportions of major cell types identified in the glioblastoma sample by scCUT&Tag versus scATAC-seq.

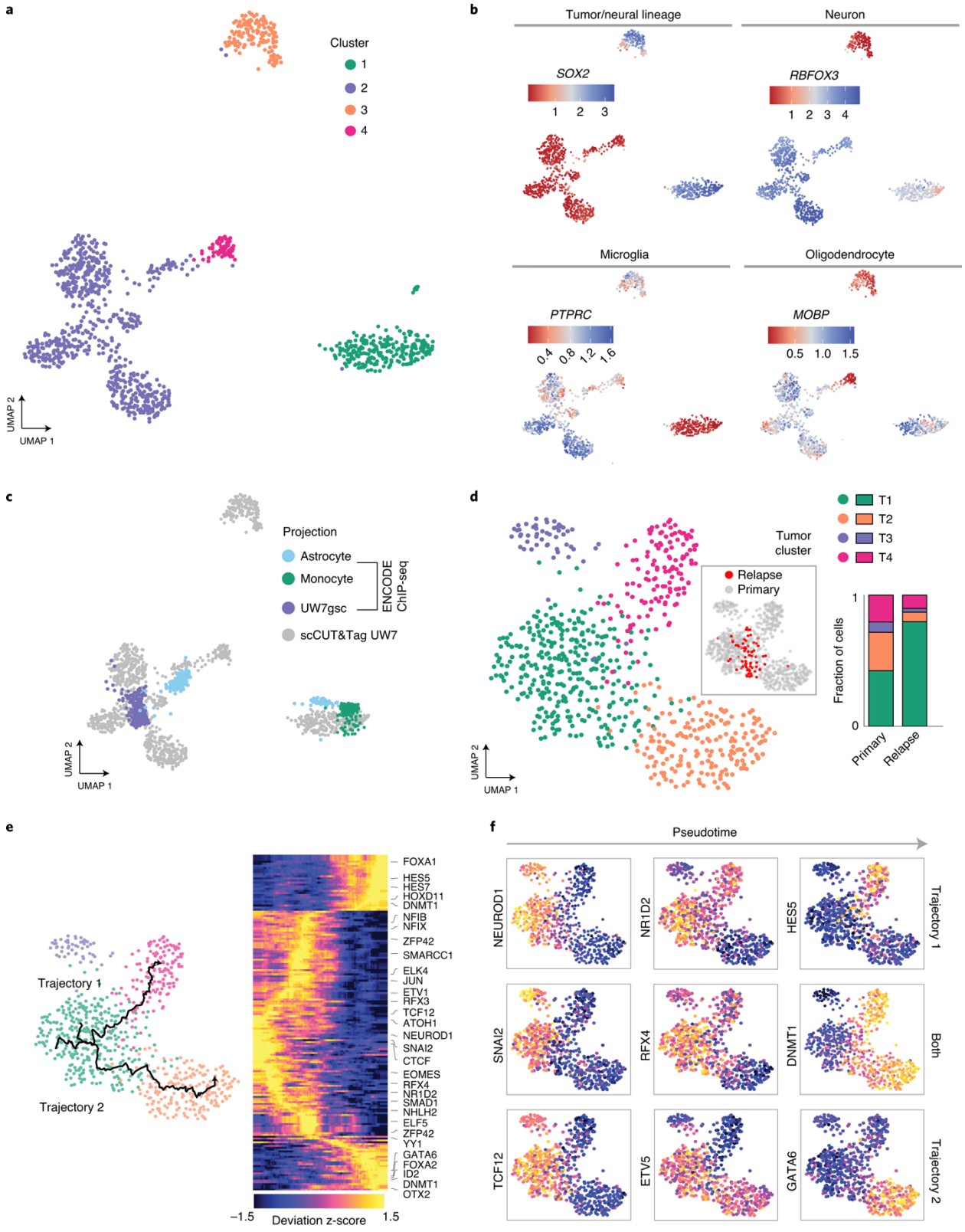

a
Cluster
1
2
3
4
UMAP 2
UMAP 1
b
Tumor/neural lineage
Neuron
SOX2
1 2 3
RBFOX3
1 2 3 4
Microglia
Oligodendrocyte
PTPRC
0.4 0.8 1.2 1.6
MOBP
0.5 1.0 1.5
c
Projection
Astrocyte
Monocyte
UW7gsc
ENCODE ChIP-seq
scCUT&Tag UW7
UMAP 2
UMAP 1
d
Tumor cluster
Relapse
Primary
T1
T2
T3
T4
Fraction of cells
1
0
Primary
Relapse
UMAP 2
UMAP 1
e
Trajectory 1
Trajectory 2
FOXA1
HES5
HES7
FOXD11
DNMT1
NFIB
NFIX
ZFP42
SMARCC1
ELK4
JUN
ETV1
RFX3
TCF12
ATOH1
NEUROD1
SNAI2
CTCF
EOMES
RFX4
NR1D2
SMAD1
NHLH2
ELF5
ZFP42
YY1
GATA6
FOXA2
ID2
DNMT1
OTX2
−1.5
Deviation z-score
1.5
f
Pseudotime
NEUROD1
NR1D2
HES5
Trajectory 1
SNAI2
RFX4
DNMT1
Both
TCF12
ETV5
GATA6
Trajectory 2

Figure 4.8: scCUT&Tag data for H3K27me3 for a human glioblastoma primary and relapse sample. a) UMAP embedding of single cells from a primary human glioblastoma based on H3K27me3 signal. b) Cluster annotation using CSS for key marker genes identifies microglia (PTPRC), neurons (RBFOX3), oligodendrocytes (MOBP) and tumor cells (SOX2). c) UMAP transform and projection of bulk ChIP-seq (monocytes and astrocytes) or bulk CUT&RUN (UW7gsc) onto patient sample. d) Left, UMAP co-embedding of tumor cells from primary and relapse sample. Inset highlights locations of cells from relapse sample. Right, bar plot demonstrating fraction of cells in each sample (Primary and Relapse) that belong to each cluster. e) Left, two pseudotime trajectories starting with cluster T1 (presumed stem-like cluster) and ending in either Cluster T4 (Trajectory 1) or Cluster T2 (Trajectory 2). Right, heat map of 132 significant motif deviations based on H3K27me3 activity within peaks from aggregated tumor cell ATAC-seq data. Motif deviations are ordered by pseudotime. f) UMAP plots for tumor cells colored by deviation scores for selected motifs. Left column shows early motifs in pseudotime that are commonly silenced, including NEUROD1, SNAI2 and TCF12. Middle column shows silenced programs that diverge according to trajectory (NR1D2 in Trajectory 1 and ETV5 in Trajectory 2) or are common across trajectories (RFX4). Right column shows silenced programs specific to terminal pseudotime for Trajectory 1 (HES5), Trajectory 2 (GATA6) or both (DNMT1).

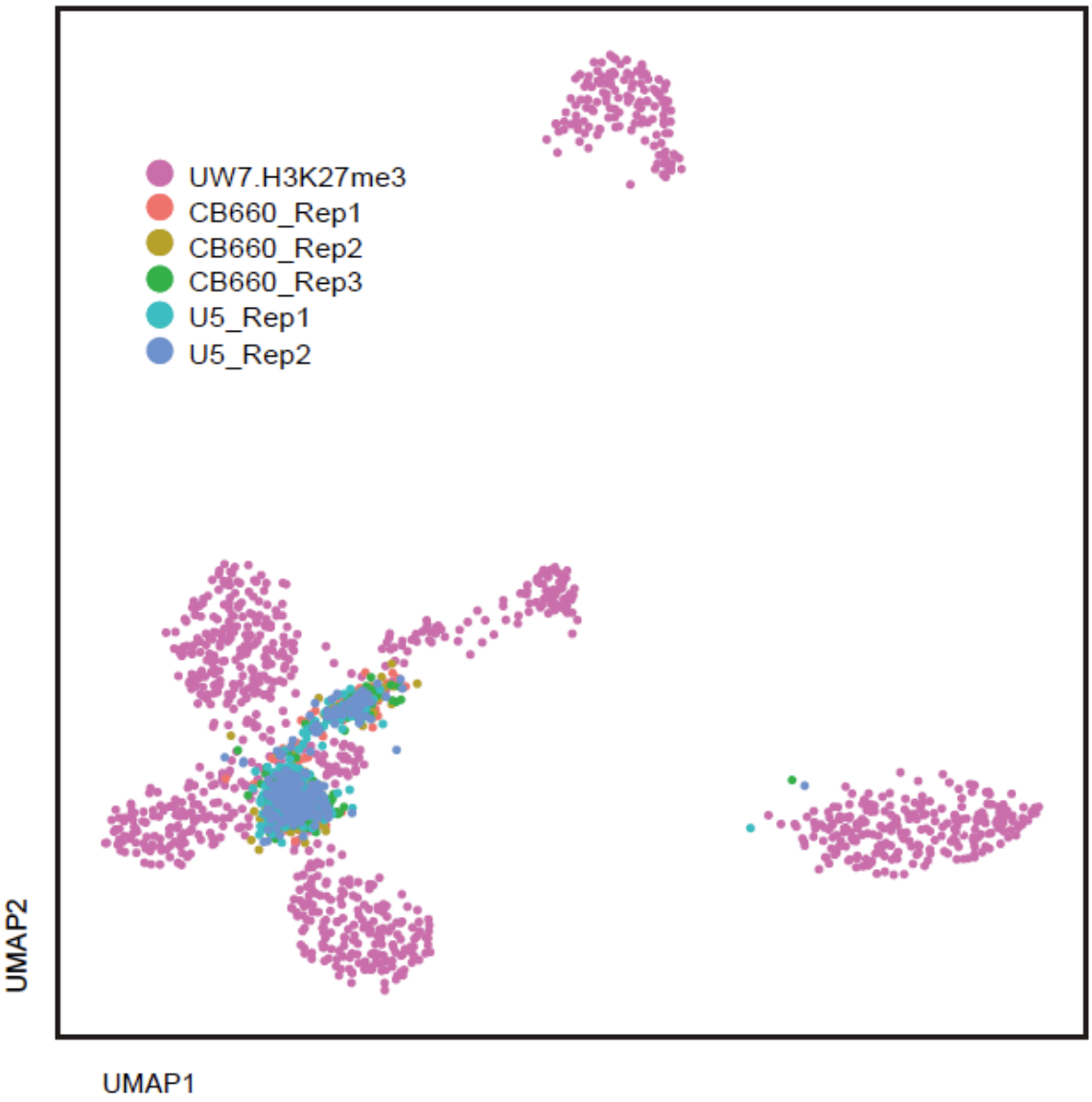


Figure 4.9: CUT&RUN data projected onto single cell patient data.
Projection of published bulk CUT&RUN data for two neural stem cell lines CB660 (three replicates) and U5 (two replicates) onto the single cell patient data. Neural stem cells localize to the center of the tumor cell cluster but also adjacent to the astrocyte cluster.

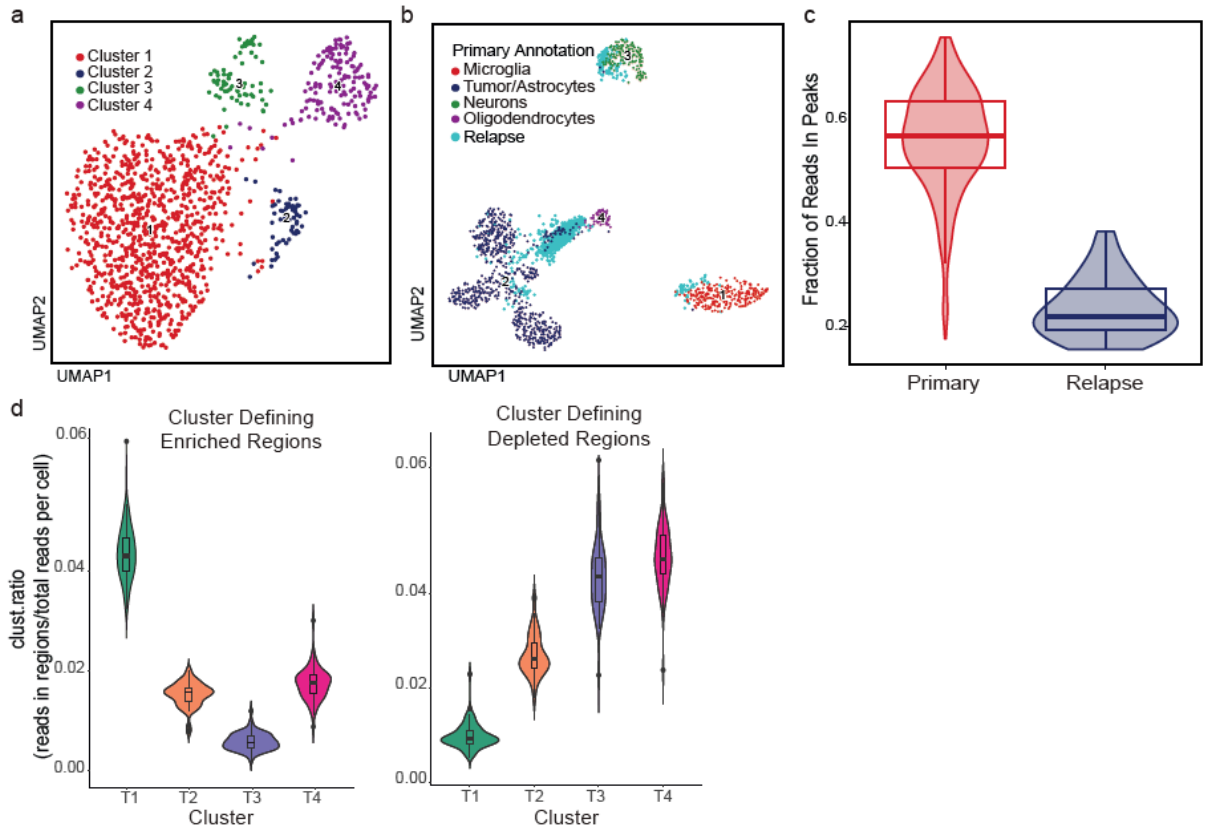


Figure 4.10: Identification of tumor cells from relapsed sample.
a) UMAP embedding of relapse sample from patient (n=1168 cells) based on H3K27me3 signal. b) Projection of single cells from relapsed sample onto primary patient sample using UMAP transform allows identification of the four clusters in the relapse data corresponding to microglia, tumor cells, astrocyes, and neurons. The majority of the cells colocalize with the astrocyte cluster from the primary tumor. 71 cells colocalized with the tumor cell cluster. c) We called peaks using HOMER with settings appropriate for broad domains (see methods) on the aggregate sequencing data for each sample and then calculated the fraction of reads in peaks (FRiP) on a per cell basis. The value for each cell (n=639 for primary and 69 for relapse sample) is plotted. The primary sample had a higher (mean 0.57±0.10 versus 0.22±0.06) FRiP value. d) Violin plots demonstrating similarity of autopsy cells (n=69) to cluster 1 from the primary tumor. Genomic regions that most significantly (FDR<0.05, Log2FC>1 or Log2FC<1) distinguished tumor clusters based on chromatin silencing score were calculated. Cluster ratio was calculated by counting reads in each autopsy cell in cluster defining regions and dividing by total number of unique fragments for that cell. Distributions were plotted for each set of clusters defining
regions (T1-T4). This was performed for regions that define each cluster by both enrichment (positive, left panel, mean=0.043±0.005, 0.015±0.002, 0.058±0.002, and 0.017±0.003 for T1, T2, T3, and T4, respectively) and depletion (negative, right panel, mean=0.01±0.003, 0.027±0.004, 0.043±0.007, and 0.048±0.007 for T1, T2, T3, and T4 respectively).

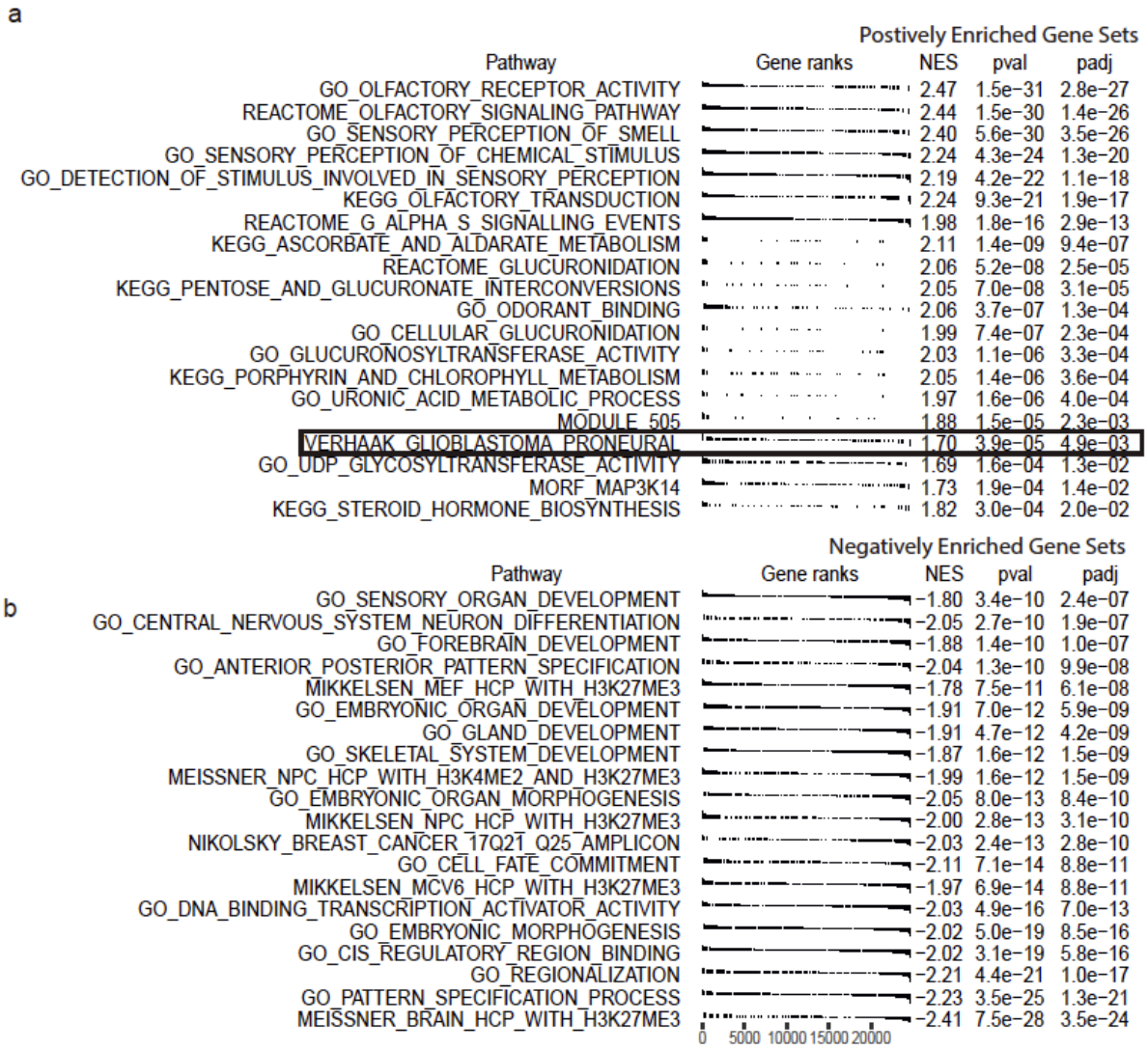


Figure 4.11: Gene set enrichment analysis of the relapse sample (cluster one).
Gene set enrichement was performed by performing a pairwise test for CSS at each gene in cluster 1 versus all other clusters. Differentially silenced genes were then rank ordered using -log10(pval)/sign(log2FC) as a metric and all MSigDB gene sets were interrogated. a) Positively enriched gene sets: genesets in this group demonstrate high PcG activity at their gene loci, and therefore represent silenced programs. The Verhaak_glioblastoma_proneural geneset is positively enriched for PcG mediated silencing. b) Negatively enriched gene sets: genesets inthe group demonstrated low PcG activity at their gene loci. The most negatively enriched
term is whole brain H3K27me3 profiles from Meissner et.al. Given that whole brain represents most differentiated cell types (neurons), it is consistent that genes in this geneset would be negatively enriched for PcG given that glioblastoma cells are not terminally differentiated.

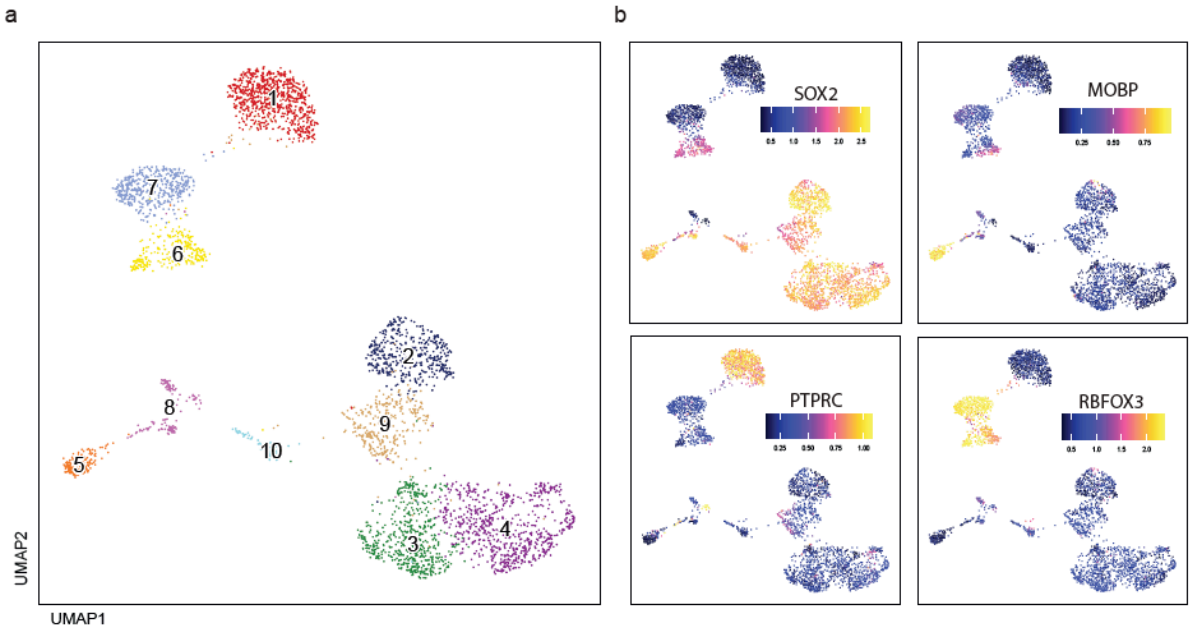


Figure 4.12: Identification of accessible chromatin peaks in scATAC-seq data.
a) UMAP embedding of single cell ATAC-seq data from patient UW7. 3541 single cells were profiled, and dimensionality reduction and clustering were performed (see methods). 10 individual clusters are identified. b) UMAP plots colored by gene activity scores for individual marker genes for each large cluster that were identified based on gene scores that were most different across the clusters. SOX2 identifies cells of neural lineage, MOBP identifies oligodendrocytes, PTPRC identifies microglia, and RBFOX3 identifies neurons. The largest subcluster of the SOX2 positive group was identified as tumor cells based on top differentially expressed genes based on gene activity scores (PTPRZ1, PDGFRA, EGFR). Cells in the tumor cell clusters were aggregated and peaks were called using MACS2, identifying accessible enhancers and promoters which were used for analysis of motif silencing in the H3K27me3 dataset.

Chapter 5

Automated CUT&Tag profiling of chromatin heterogeneity in mixed-lineage leukemia

Modified from an article originally published in *Nature Genetics:*
Derek H. Janssens, Michael P. Meers, Steven J. Wu, Ekaterina Babaeva, Soheil Meshinchi, Jay F. Sarthy, Kami Ahmad & Steven Henikoff. *Nature Genetics,* volume 53, pages 1586–1596 (2021)

Abstract

Acute myeloid and lymphoid leukemias often harbor chromosomal translocations involving the KMT2A gene, encoding the KMT2A lysine methyltransferase (also known as mixed-lineage leukemia-1), and produce in-frame fusions of KMT2A to other chromatin-regulatory proteins. Here we map fusion-specific targets across the genome for diverse KMT2A oncofusion proteins in cell lines and patient samples. By modifying CUT&Tag chromatin profiling for full automation, we identify common and tumor-subtype-specific sites of aberrant chromatin regulation induced by KMT2A oncofusion proteins. A subset of KMT2A oncofusion-binding sites are marked by bivalent (H3K4me3 and H3K27me3) chromatin signatures, and single-cell CUT&Tag profiling reveals that these sites display cell-to-cell heterogeneity suggestive of lineage plasticity.

Introduction

Ten percent of acute leukemias harbor chromosomal translocations involving the KMT2A gene encoding lysine methyltransferase 2A (also referred to as mixed-lineage leukemia-1) (Winters & Bernt, 2017). In its normal role, KMT2A catalyzes methylation of the Lys4 residue of the histone H3 nucleosome tail (H3K4) and is required for fetal and adult hematopoiesis (Antunes & Ottersbach, 2020). The N-terminal portion of KMT2A contains a low-complexity domain that mediates protein–protein interactions, an AT-hook/CXXC domain that binds DNA and multiple chromatin-interacting domains (PHD domains and a bromo domain), whereas the C-terminal portion contains a transactivation domain that interacts with histone acetyltransferases and a SET domain that catalyzes histone H3K4 methylation (Dou, 2005; Ernst et al., 2001; Milne, 2002; Zeleznik-Le et al., 1994). The KMT2A precursor protein is cleaved to a 320-kDa N-terminal fragment (KMT2A-N) and a 180-kDa C-terminal fragment (KMT2A-C) that form a stable dimer (Hsieh, Cheng, et al., 2003; Hsieh, Ernst, et al., 2003).

KMT2A contributes to leukemogenesis through oncogenic chromosomal rearrangements involving the DNA-binding domain in the N-terminal portion of KMT2A with a diverse array of other chromatin-regulatory proteins (Hu & Shilatifard, 2016; Slany, 2016). Although more than 80 translocation partners have been identified in KMT2A-rearranged (KMT2Ar) leukemias, fusions involving the AF9, ENL, ELL, AF4 and AF10 transcriptional elongation factors account for the majority of cases (Hu & Shilatifard, 2016; Winters & Bernt, 2017). These fusion partners regulate RNA polymerase II elongation (ELL and

AF4), recruit the DOT1L H3K79 histone methyltransferase (AF10), or both (AF9 and ENL) (C. Lin, 2010, p. 4; Monroe, 2011; Okada, 2005; D. T. Zeisig, 2005). Additionally, ENL and AF9 interact with the CBX8 chromobox protein to neutralize the Polycomb repressive complex 1 (PRC1) gene-silencing complex (Maethner, 2013; Tan, 2011).

Previous work has suggested that KMT2A fusion proteins bind different genomic loci depending on the fusion partner to drive different leukemia subtypes (S. Lin, 2016; Prange, 2017). For example, AF4 fusions are more common in acute lymphoid leukemia (ALL), and AF9 fusions are associated with acute myeloid leukemia (AML)1. In addition, KMT2A rearrangements are also prevalent in mixed-phenotype acute leukemia (MPAL), and numerous examples of KMT2Ar leukemias that interconvert between lineage types have been documented (Alexander, 2018; Gardner, 2016; S. Lin, 2016; Rayes et al., 2016). However, because methods for efficiently and reliably profiling KMT2A fusion-binding sites in scarce input patient samples are lacking, the relationship among KMT2A fusions, chromatin structure and lineage plasticity has been challenging to fully characterize. Here we establish a chromatin profiling platform that efficiently profiles oncogenic fusion proteins, transcription-associated complexes and histone modifications in cell lines and patient samples. By integrating these results with findings from related single-cell methods, we characterize the regulatory dynamics of KMT2Ar leukemias. We identify groups of target genes for fusion oncoproteins that show divergent patterns of active and repressive chromatin within the same sample. These patterns suggest that KMT2A fusion proteins activate distinct oncogenic networks within different cells of the same tumor and may explain the lineage plasticity associated with KMT2Ar leukemia.

Results

Mapping the binding sites of diverse KMT2A fusion proteins

Characterizing the chromatin localization of oncogenic fusion proteins has often been limited by the inability of chromatin immunoprecipitation and sequencing (ChIP–seq) to be used with small amounts of patient samples. To efficiently compare the binding sites for wild-type KMT2A and fusion proteins, we applied AutoCUT&RUN (Janssens, 2018) across a panel of four KMT2Ar leukemia cell lines and eight primary KMT2Ar patient samples sorted for CD45+ blasts. This collection spans the spectrum of KMT2Ar leukemia subtypes with diverse KMT2A translocations that create oncogenic fusion proteins with the transcriptional elongation factors AF4 (SEM, RS4;11, 1° ALL-1, 1° MPAL-2), AF9 (1° AML-3, 1° MPAL-1), ENL (KOPN-8, 1° AML-2), AF6 (ML-2) and AF10 (1° AML-4, 1° AML-5) as well as a relatively rare fusion to the cytoplasmic GTPase SEPT6 (1° AML-1) (Supplementary Table 1). With the exception of ML-2, an AML-derived cell line, these samples also contained a wild-type copy of the KMT2A locus. For comparison, we also profiled KMT2A localization in untransformed human CD34+ hematopoietic stem and progenitor cells (HSPCs), in H1 human embryonic stem cells and in the K562 leukemia cell line, each of which lacks KMT2A translocations. Antibodies to the C-terminal portion recognized only wild-type KMT2A-C, while antibodies to the N-terminal portion recognized both wild-type KMT2A-N and the fusion

proteins (Figure 5.1a). Therefore, binding sites unique to oncogenic fusion proteins could be identified by comparing the chromatin profiles obtained with antibodies specific to C-terminal and N-terminal KMT2A. We used AutoCUT&RUN to profile replicate samples with two different antibodies to the N terminus and two different antibodies to the C terminus of KMT2A, and correlation analysis of the sequencing results showed high reproducibility ($r = 0.78 \pm 0.19$; Figure 5.2a).

To evaluate our KMT2A dual-antibody approach, we first compared the KMT2A N-terminal and C-terminal profiles between our KMT2A wild-type and KMT2Ar samples. As expected, in H1 and K562 cells and CD34+ HSPCs, KMT2A-N and KMT2A-C showed nearly identical patterns of enrichment across the genome ($r = 0.82 \pm 0.08$; Figure 5.1b and Figure 5.2a,b). Strikingly, in H1 cells, KMT2A binding was generally focused in narrow peaks directly over TSSs, whereas in K562 cells and CD34+ progenitors additional regions showed wide peaks of both KMT2A-N and KMT2A-C extending from TSSs across gene bodies. Many of the genes with a wide KMT2A distribution in CD34+ progenitors (for example, HOXA9, RUNX2, MEIS1, MEF2C) are master regulators of hematopoietic cell fate (Figure 5.1c) and have previously been defined as KMT2A fusion oncoprotein targets in leukemias (Guenther, 2008; Prange, 2017). In the KMT2Ar leukemia samples, the correlation between the KMT2A N-terminal and C-terminal profiles was significantly lower than in the control samples ($r = 0.53 \pm 0.21$; Figure 5.2a,c), and many of the wide KMT2A-bound regions showed an enriched KMT2A N-terminal signal relative to the KMT2A C terminus (Figure 5.1b,c). To systematically define fusion protein-binding sites across our collection of samples, we used Gaussian mixture modeling to partition KMT2A peaks into two different distributions based on both the width of the KMT2A peaks and the enrichment-normalized ratio of KMT2A-N to KMT2A-C signal (KMT2A N/C score; c and Figure 5.3). In CD34+ HSPCs, 131 of 6,336 KMT2A-bound sites were called as wide peaks (mean = 8.2 ± 4 kb) and a two-component Gaussian mixture model failed to partition the KMT2A-bound sites based on N/C score (Figure 5.1c,d and Figure 5.3a–e), suggesting that there is similar enrichment of KMT2A-N and KMT2A-C proteins, consistent with wild-type KMT2A binding. In comparison, in the SEM cell line encoding a KMT2A–AF4 fusion, 195 of 8,259 KMT2A-bound regions were called as wide (mean = 13.1 ± 10 kb), and about half of these wide peaks (91/195) were enriched for KMT2A-N relative to KMT2A-C, which we interpret as fusion oncoprotein-binding sites (Figure 5.1c,d and Figure 5.3f–j). In line with this interpretation, 61 of 91 of the oncoprotein target genes we identified in SEM cells overlapped with KMT2A–AF4 target genes that were previously identified in SEM cells using ChIP–seq23 (Figure 5.3k). The ML-2 cell line has a deletion of the wild-type KMT2A allele and harbors only a KMT2A–AF6 fusion oncoprotein. As expected, the majority of KMT2A-bound sites had a high KMT2A N/C score (144/211; Figure 5.1c and Figure 5.3l). This persistent localization of the KMT2A–AF6 fusion oncoprotein to chromatin in ML-2 cells demonstrates that binding of the oncoprotein is not dependent on wild-type KMT2A.

Next, we examined how the distribution of KMT2A differed between wild-type and oncofusion proteins. In CD34+ HSPCs, 81% of wide peaks overlapped a gene TSS, whereas in KMT2Ar samples significantly fewer fusion oncoprotein peaks overlapped a TSS (30%), and significantly more (50%)

overlapped a gene body (Figure 5.1e). In comparison, in the control H1 human embryonic stem cell line, only 15 of 17,000 KMT2A peaks were called as wide and these peaks were significantly less enriched on gene TSSs than the wide KMT2A peaks we identified in CD34+ HSPCs and significantly less enriched in gene bodies than the oncoprotein-binding sites in KMT2Ar leukemia samples (Figure 5.1e and Figure 5.3m). This pattern of KMT2A fusion oncoproteins spreading across target gene bodies is consistent with previous reports (Guenther, 2008; Kerry, 2017). By comparing enrichment of the KMT2A N terminus and C terminus across fusion oncoprotein-binding sites in all KMT2Ar samples, we found that in all cases the N terminus was significantly more enriched than the C terminus (Figure 5.1f), and in all cases the fold difference between the N-terminal and C-terminal signal at oncoprotein-bound regions in KMT2Ar samples was greater than in the wide KMT2A-bound regions in CD34+ HSPCs (Figure 5.4a). Taking these findings together, we conclude that our KMT2A N-terminal versus C-terminal antibody multiplexing approach identifies regions bound by diverse KMT2A fusion oncoproteins.

We then compared oncoprotein target sites among leukemias with different KMT2A fusions. We found that 81 of 440 (~18%) of all fusion oncoprotein target genes were shared by five or more of the KMT2Ar leukemia samples we profiled, representing 12% of the total sequence space occupied by the fusion proteins (Figure 5.4b,c). As expected, the group of genes we identified as the most frequent KMT2A fusion targets across our collection of samples included several genes that are known to be required for KMT2Ar leukemia, including HOXA9, MEIS1, MEF2C, MBNL1 and JMJD1C (Cante-Barrett et al., 2014; Itskovich, 2020; Izaguirre-Carbonell, 2019; Wong et al., 2007; B. B. Zeisig, 2004) (Figure 5.4c). PCA of KMT2A N/C scores across all oncoprotein-binding sites indicated that both the specific fusion partner and the myeloid versus lymphoid lineage bias of the tumor may influence tumor-specific localization of the oncofusion protein (Figure 5.1g). For example, all KMT2A–AF4 samples clustered together in the PCA plot and grouped with a sample from a patient with ALL and a sample from a patient with MPAL. By contrast, the ALL cell line KOPN-8 carries a KMT2A–ENL fusion protein and partitioned away from KMT2A–AF4-bearing leukemias. Primary AML samples bearing KMT2A–AF9, KMT2A–AF10 and KMT2A–ENL fusions formed a second cluster, apart from the KMT2A–SEPT6-containing primary AML and the primary KMT2A–AF9-bearing MPAL sample. Thus, tumors bearing KMT2A–AF4 fusions share a distinct binding profile, but other oncofusion proteins such as KMT2A–ENL and KMT2A–AF9 also appear to be influenced by the lineage bias of the tumor.

Chromatin landscape of KMT2Ar leukemia samples

To economically characterize the global chromatin landscape of tumors at a scale that could be generally applied to patient samples, we developed AutoCUT&Tag, a modification of our previous AutoCUT&RUN robotic platform (Janssens, 2018). CUT&Tag takes advantage of the high efficiency and low background of antibody-tethered Tn5 tagmentation-based chromatin profiling relative to previous methods, such as ChIP–seq and CUT&RUN (Kaya-Okur, 2019). The standard CUT&Tag protocol requires DNA extraction before library enrichment by PCR. However, we recently developed conditions

for DNA release and PCR enrichment without extraction (CUT&Tag-direct) (*Kaya-Okur, H. S., Janssens, D. H., Henikoff, J. G., Ahmad, K. & Henikoff, S. Efficient Low-Cost Chromatin Profiling with CUT&Tag. Nat. Protoc.15, 3264–3283 (2020).*, n.d.). In this improved protocol, a low concentration of SDS is used to displace bound Tn5 from tagmented DNA, and subsequent addition of the non-ionic detergent Triton X-100 quenches the SDS to allow for efficient PCR. This streamlined protocol makes CUT&Tag compatible with robotic handling of samples in a 96-well plate format and generates profiles with data quality comparable to that produced by benchtop CUT&Tag ($r = 0.79 \pm 0.093$; Figure 5.5a–c).

To define the chromatin features around KMT2A fusion-binding sites, we used AutoCUT&Tag to profile the active chromatin modifications H3K4me1, H3K4me3, H3K36me3, H3K27ac and H4K16ac as well as initiating RNA polymerase II marked by Ser5 phosphorylation of the C-terminal domain (RNAP2S5p). In addition, we profiled the silencing histone modifications H3K27me3 and H3K9me3. Together, these eight modifications distinguish active promoters, enhancers, transcribed regions, developmentally silenced chromatin and constitutively silenced chromatin (J. Zhang, 2020) and provide a straightforward picture of the regulatory status of a genome (Figure 5.6a). Replicate profiles for each mark in control CD34+ samples and KMT2Ar leukemia samples were very similar and were merged for further analysis (H3K27me3, $r = 0.93 \pm 0.051$; H3K4me3, $r = 0.96 \pm 0.015$; H3K4me1, $r = 0.90 \pm 0.037$; H3K9me3, $r = 0.83 \pm 0.060$; H3K27ac, $r = 0.80 \pm 0.077$; H3K36me3, $r = 0.95 \pm 0.021$; H4K16ac, $r = 0.97 \pm 0.012$; RNAP2S5p, $r = 0.77 \pm 0.107$) (Figure 5.5d).

We first compared the chromatin features associated with sites bound by wild-type KMT2A to sites bound by KMT2A oncofusion proteins across all samples. In line with localization of the KMT2A fusion proteins to actively transcribed genes, we found that the active promoter marks H3K4me3, RNAP2S5p and H3K27ac were all present at oncofusion protein-binding sites (Figure 5.7a–c). H3K4me3 was also enriched at some promoters in the ML-2 cell line (for example, LPO and LYZ in Figure 5.6b), which lacks the KMT2A methyltransferase domain, indicating that another H3K4me3 methyltransferase is responsible. In comparison to sample-matched wild-type KMT2A-bound sites, H3K27ac was enriched at oncofusion protein-binding sites in all samples, but this difference was statistically significant in only 6 of 11 samples (SEM, KOPN-8, 1° AML-1, 1° AML-2, 1° AML-3 and 1° AML-5; Figure 5.7b). The H3K4me3 mark was significantly enriched at oncofusion protein-binding sites in five of the samples (SEM, RS4;11, 1° AML-1, 1° AML-2 and 1° MPAL-2) and significantly depleted in five of the other samples (1° ALL-1, 1° AML-3, 1° AML-4, 1° AML-5 and 1° MPAL-1; Figure 5.7c). Oncofusion protein-binding sites lacked H3K27me3 and H3K9me3 (Figure 5.7d,e) but were enriched in H3K4me1 and H3K36me3, both of which mark transcribed gene bodies, and this enrichment was significant in 9 of 11 and 6 of 11 of the KMT2Ar leukemia samples, respectively (Figure 5.7f,g). Enrichment of these marks is expected with mistargeting of KMT2A fusions to gene bodies (Chan & Chen, 2019).

Histone modification profiling holds the potential to reveal similarities and distinctions between leukemias by reporting their transcriptional status. For example, H3K4me3 reports gene promoter activity and was enriched at marker genes that matched the immunophenotypic characterization of each

leukemia (Figure 5.6b). To determine how the global distribution of these marks varied between KMT2Ar leukemia samples, we first identified regions enriched for each modification in our collection of KMT2Ar leukemia samples as well as CD34+ HSPCs using the SEACR peak-calling method (Meers, Tenenbaum, et al., 2019) and performed PCA to cluster samples according to their modification-specific similarities. Overall, active chromatin features marked by H3K4me1, H3K4me3, H3K36me3, H4K16ac or RNAP2S5p clustered samples according to their ALL, AML or MPAL lineage designation (Figure 5.6c–e and Figure 5.8a,b), suggesting that a similar repertoire of active genes is used in each leukemia subtype. By contrast, PCA based on H3K27ac or H3K27me3 CUT&Tag profiles partitioned samples into groups largely unrelated to leukemia subtype (Figure 5.6f and Figure 5.8c), and only the 1° AML-1 sample was distinguished by H3K9me3 (Figure 5.8d). H3K27me3 is an epigenetically inherited histone modification that is linked to developmental progression as cells determine their identity. Thus, these distinct H3K27me3 leukemia landscapes may be related to hematopoietic transitions that are defective in each tumor.

We next examined the lineage-specific variation in gene and regulatory element usage as indicated by the global chromatin landscape of each of the marks we profiled by performing t-SNE of these elements followed by density peak clustering (Yoshida, 2019). This analysis revealed that H3K4me1-marked regions were highly variable between lineage subtypes, with a substantial fraction of elements marked specifically in AML samples falling to one side of the t-SNE plot (8,221/56,267), ALL-specific elements partitioned to the other side of the plot (8,466/56,267) and CD34+ HSPC elements grouped in the middle (7,141/56,267) (Figure 5.6g and Figure 5.8e,f). A fraction of both the AML- and ALL-specific elements were also marked by H3K4me1 in CD34+ cells and the primary MPAL samples we profiled (Figure 5.6g and Figure 5.8e,f). This regulatory overlap implies that MPAL leukemias share features with both ALL and AML and that KMT2Ar leukemia samples maintain H3K4me1 at regulatory elements used during normal hematopoiesis.

While only about half of the H3K4me1-marked elements were similarly labeled across all samples (~50%, 25,973/56,267), a much larger fraction of H3K4me3 (~75%, 10,958/13,998) and H3K36me3 (~85%, 22,858/26,759) peaks were common across leukemia subtypes, indicating that these subtypes largely share gene expression repertoires (Figure 5.6h,i and Figure 5.8g–j). Grouping H3K4me3-marked promoter regions by t-SNE also partitioned 64 AML- and 508 ALL-specific elements to opposite sides of the t-SNE graph and identified 1,918 elements that were shared with the MPAL samples and CD34+ HSPCs (Figure 5.6h and Figure 5.8g,h); however, as compared to H3K4me1, where we identified 23,828 lineage-specific peaks among the 56,267 total peaks (~40%), a smaller proportion of H3K4me3-marked features showed any lineage specificity (2,490/13,998 peaks, or ~18%). This is consistent with previous reports that regulatory elements marked by H3K4me1 generally show more cell type specificity than promoter elements marked by H3K4me3 (Gorkin, 2020; Heintzman et al., 2009).

Similarly to the t-SNE analysis of H3K36me3-marked regions, t-SNE analysis of H3K27ac-, H4K16ac- and H3K9me3-marked regions did not partition the genome by lineage identity (Figure 5.9a–c).

By contrast, both RNAP2S5p and H3K27me3 peaks showed diversity similar to that observed with H3K4me1 (Figure 5.6j and Figure 5.9d). Analysis by t-SNE with H3K27me3 did not partition elements according to lineage subtype (Figure 5.6j). Rather, AML and ALL samples had a significantly greater proportion of the genome that was marked by H3K27me3 than CD34+ cells (Figure 5.6j and Figure 5.9e), suggesting that these tumor types are more differentiated. In line with this interpretation, MPAL samples had significantly fewer regions marked by H3K27me3 than the ALL or AML samples and were considered to have a higher degree of lineage plasticity (Figure 5.6j and Figure 5.9e). We conclude that high-throughput CUT&Tag profiling provides a powerful tool to characterize KMT2Ar leukemias and that profiling the developmentally repressed genome reveals tumor-specific differences that are not apparent by profiling the active genome.

Bivalent chromatin at KMT2A fusion protein target sites

In addition to marking promoters that are engaged in active transcription, H3K4me3 is present at a limited subset of transcriptionally repressed 'bivalent' (that is, 'poised') promoters that are also marked by H3K27me3 (Bernstein et al., 2006; Cui, 2009). In our collection of leukemia samples, we observed both H3K4me3 and H3K27me3 at some promoters that were called as KMT2A fusion protein targets (Figure 5.6a, left). Additionally, we observed genes that were bound by the oncofusion protein in the majority of KMT2Ar leukemia samples but were not called as targets in specific samples; we termed this group 'missing targets' (Figure 5.6a, right). To systematically define bivalent promoters within our collection of samples, we quantified the abundance of H3K4me3 and H3K27me3 in 2-kb windows centered on gene TSSs of marked and unmarked promoters for each modification. By intersecting these groups, we identified approximately 2,000–5,000 bivalent promoters in each of the KMT2Ar leukemia samples (Figure 5.6k and Supplementary Table 3). Interestingly, we found that approximately 33% (129/396) of promoters for missing targets were called as bivalent, whereas approximately 24% (267/1,097) of KMT2A fusion target promoters were bivalent and only 14% of wild-type KMT2A target promoters were bivalent (Figure 5.6l). Thus, oncofusion protein target promoters are enriched for a bivalent chromatin signature, suggesting that expression of these genes may fluctuate among cells within a sample.

Cell-to-cell chromatin heterogeneity at KMT2A fusion targets

To test whether the bivalent chromatin signature at KMT2A fusion target promoters is due to heterogeneity among cells, we performed single-cell CUT&Tag on four KMT2Ar cell lines and four primary KMT2Ar leukemia samples. Antibody binding and pA-Tn5 tethering were performed on bulk samples, and individual cells were then arrayed in microwells on the ICELL8 platform for barcoded PCR library enrichment (Kaya-Okur, 2019). We optimized the median number of unique reads per cell while maintaining a high fraction of reads in peaks (FRiP) on the ICELL8 by varying the amount of SDS detergent used to release Tn5 after tagmentation and the amount of Triton X-100 used to quench SDS

before PCR (Figure 5.10a,b). Using this approach, we profiled 1,137–3,611 cells for the H3K4me3, H3K27me3 and H3K36me3 histone modifications. After excluding cells with fewer than 300 fragments, single-cell CUT&Tag for H3K4me3, H3K27me3 and H3K36me3 yielded a median of 4,972, 13,025 and 3,962 unique reads per cell, respectively (Figure 5.10c). As a second quality-control step, we called peaks on the aggregate data of all cells profiled for each mark and removed cells that had a FRiP value below the normal distribution (Figure 5.10d,e). Profiles for each single cell were then split into 5-kb bins tiled across the genome, and cells were projected in UMAP space on the basis of this binning (Figure 5.11a–c). Encouragingly, cells taken from the same leukemia sample and profiled in different experiments were clustered together in UMAP space, indicating that the data quality was consistent between batches (Figure 5.10f–h). This approach resolved clusters for samples based on H3K4me3 or H3K27me3 profiling but not H3K36me3 profiling (Figure 5.11a–c). This implies that the leukemia samples differ in both sets of active promoters and silenced regions.

To examine intratumoral heterogeneity in the H3K4me3 and H3K27me3 signals, we first used the archR single-cell software package (Dijk, 2018; Granja, 2021) to calculate imputed gene scores for all genes according to the UMAP projection of all cells. We then determined the normalized dispersion of the imputed scores in cells from the same sample (Figure 5.11d,e). Strikingly, bivalent missing targets showed significantly higher H3K4me3 dispersion in the SEM, KOPN-8, 1° ALL-1, 1° AML-2, 1° MPAL-1 and 1° MPAL-2 samples than tumor-matched control genes (Figure 5.11f). This implies that levels of the H3K4me3 active promoter mark in these genes vary among cells within KMT2Ar leukemias.

Next, we examined variation in the repressive H3K27me3 mark at bivalent oncoprotein target genes. In 1° MPAL-1 cells, the normalized dispersion of H3K27me3 was significantly higher in bivalent missing target genes, and in the 1° MPAL-2 sample the normalized H3K27me3 dispersion was higher in bivalent target genes than in tumor-matched control genes (Figure 5.11f). Some bivalent genes varied among cells for both the H3K4me3 and H3K27me3 modifications. For example, the HOXA9 gene was a missing target in 1° MPAL-1 cells (Figure 5.6a) but showed high dispersion in both the H3K4me3 and H3K27me3 signals (Figure 5.11d,e). Thus, bivalency of chromatin marks is associated with heterogeneity among cells within a sample.

Grouping bivalent target genes according to the Pearson correlation of their imputed gene scores across cells of a given leukemia sample separated two groups by either H3K4me3 or H3K27me3 profiling (Figure 5.11g,h and Figure 5.12a–g). For example, the missing target gene HOXA9 had elevated H3K4me3 scores in a small fraction of 1° MPAL-1 leukemia cells (~15%) (Figure 5.11i and Figure 5.12h). The TAPT1 gene clustered together with HOXA9 (Figure 5.11g) and, as expected, had the highest H3K4me3 scores in the same cells as HOXA9 ($r = 0.87$) (Figure 5.11i). By contrast, genes that were anticorrelated with HOXA9, such as CPEB2 ($r = -0.74$) and MEIS1 ($r = -0.42$), had the weakest H3K4me3

signal in cells where HOXA9 was active (Figure 5.11i). This suggests that there are two exclusive gene expression programs activated by KMT2A fusion oncoproteins. Furthermore, we found that the imputed H3K27me3 scores also formed inverse patterns of gene association from H3K4me3, where genes such as HOXA9 that were rarely marked by H3K4me3 in 1° MPAL-1 leukemia cells showed elevated H3K27me3 scores in the majority of tumor cells (~55%) (Figure 5.11j and Figure 5.12i). These groups of divergent KMT2A fusion oncoprotein targets may contribute to the phenotypic plasticity of KMT2Ar leukemias.

Discussion

Here we have applied high-throughput chromatin profiling to KMT2Ar leukemias to delineate fusion protein-specific targets and to identify chromatin features that are characteristic of myeloid, lymphoid and mixed-lineage leukemias. To economically profile these features, we took advantage of the high signal-to-noise ratio and low sequencing depth requirements inherent to CUT&RUN and CUT&Tag and fully automated both methods on a standard liquid handling robot. As CUT&Tag requires only thousands of cells for informative histone modifications (Henikoff et al., 2021), AutoCUT&Tag is suitable for profiling of samples for a wide range of studies, including developmental and disease studies, and screening of patient samples. The enhanced throughput and consistency of the AutoCUT&RUN and AutoCUT&Tag platforms for chromatin profiling make these technologies suitable for profiling patient specimens.

By also performing AutoCUT&RUN on KMT2A fusions and components of the Super Elongation and DotCom complexes, we have elucidated the details of mechanisms that likely contribute to the heterogeneity of these tumors. We found that the most common KMT2A fusion proteins, including KMT2A–AF4, KMT2A–AF9, KMT2A–ENL and KMT2A–AF10, all colocalize with the DOT1L and ENL proteins in gene bodies. This suggests that interaction of the C-terminal domains of AF4, AF9, ENL and AF10 with transcriptional elongation complexes likely recruits fusion proteins from the promoter into the gene body. In line with the possibility that these interactions have a pivotal role in oncogenic transformation, wild-type ENL protein is required for tumor growth in numerous KMT2Ar cell lines (Erb, 2017).

Using AutoCUT&Tag to profile histone modifications in leukemia samples, we identified frequent KMT2A fusion oncoprotein sites with bivalent chromatin features. At some sites, bivalent chromatin features correlated with heterogeneity among cells of the same tumor, which suggests that the heterogeneity in gene expression seen in populations of mixed-lineage leukemia cells is rooted in chromatin dynamics. We identify a group of KMT2A oncoprotein target genes that are shared in the majority of KMT2Ar leukemias but are missed in a subset of samples. In several of the KMT2Ar leukemia

samples we profiled, these missing targets were among the genes that showed the highest variation in active and repressive chromatin marks within the tumor, suggesting that these missing targets may be bound and activated by the oncoprotein in a limited subset of cells within the tumor, causing them to fall below the levels necessary for detection in our bulk KMT2A profiling assays. This heterogeneity we observed at KMT2A oncoprotein target genes has implications for how resistance to therapies may develop, if only a subset of cells are susceptible to specific anticancer agents.

Heterogeneity in leukemias may arise if an early cancerous cell divides and differentiates into two related cell types. Alternatively, certain leukemias may sporadically switch between cell types (Forgione et al., 2020). Our single-cell profiling reveals that some leukemias display both active and repressive chromatin states at KMT2A fusion target loci that differ among individual cells. Kinetic analysis of chromatin dynamics within cell populations will be needed to determine whether bivalency reflects differentiation or sporadic switching, with implications for therapeutic strategies to limit relapse.

Methods

Patients

All patient samples were obtained by St. Jude Children's Research Hospital or member COG institutions in accordance with the Declaration of Helsinki after written consent from the parents/guardians of minors upon enrolling in the trial. The studies were overseen by the institutional review boards at Fred Hutchinson Cancer Research Center (IR protocol 9950) and St. Jude Children's Research Hospital. Patients did not receive compensation for participation in this study.

Cell culture

Human K562 cells were purchased from ATCC (CCL-243) and cultured according to the supplier's protocol. H1 human embryonic stem cells were obtained from WiCell (WA01-lot# WB35186) and cultured in plates coated with Matrigel (Corning) in mTeSR1 Basal Media (STEMCELL Technologies, 85851) containing mTeSR1 Supplement (STEMCELL Technologies, 85852). The KMT2Ar cell lines ML-2, KOPN-8, RS4;11 and SEM were obtained from the Bleakley laboratory at the Fred Hutchinson Cancer Research Center. The SEM cell line was cultured in IMDM (ThermoFisher, 12440061) supplemented with 10% FBS. The ML-2, KOPN-8 and RS4;11 cell lines were cultured in RPMI 1640 with glutamine and HEPES (ThermoFisher, 72400047) supplemented with 10% FBS. All cell lines were maintained in a cell culture incubator (Sanyo, MCO-19AIC) with standard settings (37 °C with 5% CO2).

Drug treatment

Ten thousand SEM, RS4;11 and KOPN-8 cells were plated in 90 µl of the appropriate medium (see above) in a 96-well cell culture plate. Serial dilutions of either the DOT1L inhibitor EPZ-5676 (MedChem Express, HY-15593) or the Menin inhibitor VTP50469 (MedChem Express, HY-114162) were

prepared in DMSO and then diluted in primary medium to control for the concentration of DMSO across all conditions. Ten microliters of the diluted inhibitors was then added to cell culture suspensions followed by mixing. Cells were grown for 3 or 4 d, at which point viability was measured using a CellTiter-Glo assay (Promega, G9241) read out on a standard luminometer. For chromatin profiling experiments, SEM, RS4;11 and KOPN-8 cells were plated at the same density (10,000 cells per 100 µl) in 20 ml of medium containing 30 µM EPZ-5676, 30 µM VTP50469 or DMSO alone. After 3 d in culture, the cells were harvested and prepared for either AutoCUT&RUN or AutoCUT&Tag processing.

Primary patient samples

Diagnosis of acute leukemia were made by hematopathologists at the respective institutions based on review of histological, cytogenetic, flow cytometry and molecular studies of bone marrow biopsy samples and aspirates in accordance with World Health Organization guidelines52. Whole blood from patients with bone marrow blast percentages above 88% was subjected to Ficoll centrifugation to remove red blood cells and neutrophils. Ten million mononuclear cells were resuspended in FBS with 10% DMSO and slowly frozen in a Mr. Frosty isopropanol cannister for 24 h before being transferred to a liquid nitrogen tank. Cryopreserved leukemia blasts for 1° MPAL-1 (sample ID: SJMPAL012424_D1, alias TB-11-3295) and 1° ALL-1 (sample ID: SJALL048347_D1, alias TB-13-0939) were obtained from St. Jude Children's Research Hospital in accordance with institutional regulatory practices. Cryopreserved leukemia blasts for 1° AML-1 (sample ID: A40725), 1° AML-2 (sample ID: A67194), 1° AML-3 (sample ID: A107909), 1° AML-4 (sample ID: A38481), 1° AML-5 (sample ID: A109016) and 1° MPAL-2 (sample ID: A58548) were obtained from the Meshinchi laboratory at the Fred Hutchinson Cancer Research Center. The KMT2A fusion present in each sample was determined by whole-genome and targeted capture sequencing as previously described (A. K. Andersson, 2015; Bolouri, 2018). Cryopreserved CD34+ HSPCs from a single granulocyte colony-stimulating factor (G-CSF)-mobilized donor, enriched using a Miltenyi CliniMacs device without expansion in culture, were obtained from the Fred Hutchinson Cooperative Centers of Excellence in Hematology Core in accordance with institutional regulatory practices.

Antibodies

For profiling wild-type and oncogenic KMT2A proteins, we used two antibodies targeting the KMT2A N terminus (mouse monoclonal anti-KMT2A (1:100; Millipore, clone N4.4, 05-764) referred to as KMT2A-N1 and rabbit monoclonal anti-KMT2A (1:100; Cell Signaling Technology, clone D2M7U, 14689) referred to as KMT2A-N2) as well as two antibodies targeting the KMT2A C terminus (mouse monoclonal anti-KMT2A (1:100; Millipore, clone 9-12, 05-765) referred to as KMT2A-C1 and mouse monoclonal anti-KMT2A (1:100; Santa Cruz, clone H-10, sc-374392) referred to as KMT2A-C2). Because pA-MNase does not bind efficiently to many mouse antibodies, we used rabbit anti-mouse IgG (1:100; Abcam, ab46540) as an adaptor; this antibody was also used in the absence of a primary antibody as an IgG negative

control. For profiling Menin via AutoCUT&RUN, we used rabbit polyclonal anti-Menin (1:50; Bethyl, A300-105A). For profiling Super Elongation and DotCom components via manual and automated CUT&Tag, we used rabbit monoclonal anti-ENL (1:50; Cell Signaling Technology, clone D9M4B, 14893) and rabbit monoclonal anti-DOT1L (1:50; Cell Signaling Technology, clone D4O2T, 90878). For profiling histone marks via manual and automated CUT&Tag, as well as single-cell CUT&Tag, we used rabbit oligoclonal anti-H3K4me1 (1:100; Thermo, 710795), rabbit polyclonal anti-H3K4me3 (1:100 for bulk profiling or 1:10 for single-cell experiments; Active Motif, 39159), rabbit polyclonal anti-H3K36me3 (1:100 for bulk profiling or 1:10 for single-cell experiments; Epicypher, 13-0031), rabbit monoclonal anti-H3K27me3 (1:100 for bulk profiling or 1:10 for single-cell experiments; Cell Signaling Technology, clone C36B11, 9733), rabbit polyclonal anti-H3K9me3 (1:100; Abcam, ab8898), rabbit monoclonal anti-H3K27ac (1:50; Millipore, clone RM172, MABE647), rabbit monoclonal anti-H4K16ac (1:50; Abcam, ab109463) and rabbit monoclonal anti-RNAP2S5p (1:100; Cell Signaling Technology, clone D9N5I, 13523). To increase the local concentration of pA-Tn5, all CUT&Tag reactions also included the secondary antibody guinea pig anti-rabbit IgG (1:100; antibodies-online, ABIN101961).

AutoCUT&RUN

Primary patient samples were thawed at room temperature, washed and bound to concanavalin-A (ConA) paramagnetic beads (Bangs Laboratories, BP531) for magnetic separation. Samples were then suspended in antibody binding buffer and split for incubation with antibodies specific to the KMT2A N or C terminus or IgG control antibody overnight. Sample processing was performed by the CUT&RUN core facility at the Fred Hutchinson Cancer Research Center according to the AutoCUT&RUN protocol available from the protocols.io website (https://doi.org/10.17504/protocols.io.ufeetje).

CUT&Tag

Manual CUT&Tag reactions were performed according to the CUT&Tag-direct protocol31. Briefly, nuclei were prepared by suspending cells in NE1 buffer (20 mM HEPES-KOH pH 7.9, 10 mM KCl, 0.5 mM spermidine, 0.1% Triton X-100, 20% glycerol) for 10 min on ice. Samples were then spun down and resuspended in wash buffer (20 mM HEPES pH 7.5, 150 mM NaCl, 0.5 mM spermidine, Roche Complete Protease Inhibitor EDTA-Free) and lightly cross-linked by addition of 16% formaldehyde to a final concentration of 0.1%. After 2 min, cross-linking was stopped by addition of 2.5 M glycine to a final concentration of 75 mM. Nuclei were washed and either cryopreserved in a Mr. Frosty chamber for long-term storage or bound to ConA magnetic beads for further processing. ConA-bound nuclei were suspended in antibody binding buffer (wash buffer containing 2 mM EDTA) and split into individual 0.5-ml tubes for incubation with antibody at room temperature for 1 h or 4 °C overnight. Samples were then washed to remove unbound primary antibody, resuspended in wash buffer containing the secondary antibody and incubated at 4 °C for 1 h. Samples were washed and resuspended in 300-wash buffer (wash buffer with 300 mM NaCl) containing pA-Tn5 (1:150 dilution) and incubated at 4 °C for 1 h. Samples were

then washed in 300-wash buffer and resuspended in tagmentation buffer (300-wash buffer plus 10 mM MgCl2) and incubated at 37 °C for 1 h to allow the Tn5 tagmentation reaction to go to completion. Samples were washed with TAPS wash buffer (10 mM TAPS with 0.2 mM EDTA) and resuspended in 5 µl of release solution (10 mM TAPS with 0.1% SDS). Samples were then incubated in a thermocycler with a heated lid at 58 °C for 1 h to release Tn5 and prepare tagmented chromatin for PCR. Neutralizing solution (15 µl of 0.67% Triton X-100) was added followed by 2 µl of barcoded i5 primer (10 µM), 2 µl of barcoded i7 primer (10 µM) and 25 µl of NEBNext PCR mix. Samples were then placed in a thermocycler and PCR amplification was performed using 12–14 rapid cycles. CUT&Tag libraries were cleaned with a single round of SPRIselect beads (Beckman Coulter, B23319) at a 1.3 to 1 (vol/vol) ratio of beads to sample, quantified on a TapeStation Bioanalyzer instrument and pooled for sequencing.

AutoCUT&Tag

A detailed protocol complete with program downloads has been made publicly available on protocols.io for implementing AutoCUT&Tag on a Beckman Coulter Biomek liquid handling robot (https://doi.org/10.17504/protocols.io.bgztjx6n). To facilitate adaptation of the method to other standard liquid handling modules, the complete specifications for each step in the automated procedure are outlined in the guidelines section. Briefly, nuclei were extracted, lightly cross-linked, bound to ConA beads and incubated with primary antibody as in manual CUT&Tag. Up to 96 samples were then arrayed in a 96-well PCR plate and positioned on a stationary ALP on the Beckman Coulter Biomek FX robot equipped with an ALPAQUA Magnet Plate for standard magnetic separation, an ALPAQUA LE Magnet Plate for low-volume elution and a thermal block for temperature-controlled incubation. Wash buffer and 300-wash buffer were loaded in deep-well plates and secondary antibody solution, pA-Tn5 solution, tagmentation buffer, TAPS buffer and release buffer were all loaded into V-bottom plates and were positioned on stationary ALPs in accordance with the preprogrammed AutoCUT&Tag method. AutoCUT&Tag processing was conducted over the course of 4 h. The sample plate containing ConA-bound tagmented nuclei in 10 µl of 0.1% SDS was then removed, sealed and placed on a thermocycler with a heated lid for a 1-h incubation at 58 °C. Using a reservoir and multichannel pipettor, 54 µl of 0.15% SDS neutralization solution was added to each well, followed by 4 µl of premixed i5 and i7 barcoded primers and 36 µl of premixed KAPA PCR Master Mix. The plate was then sealed and returned to a thermocycler for 14 rapid PCR cycles. Following PCR amplification, the sample plate was returned to the Biomek for one round of post-PCR cleanup on the Biomek deck setup in accordance with a preprogrammed post-PCR cleanup method, including a second 96-well plate preloaded with SPRIselect beads, a deep-well plate loaded with 80% ethanol for bead washes and two V-bottom plates preloaded with 10 mM Tris-HCl pH 8.0 for tip washes and elution. Upon completion of the 1-h cleanup, the samples were quantified using a TapeStation Bioanalyzer instrument and pooled for sequencing.

Single-cell CUT&Tag

Nuclei were extracted and lightly cross-linked using the same strategy as for manual CUT&Tag. The nuclei concentration was then quantified using a Vi-CELL analyzer (Beckman Coulter) to allow for accurate dilution to 400 nuclei per µl (see below) before dispensing into nanowells on the ICELL8. For each antibody, 10 µl of ConA beads were washed in binding buffer (20 mM HEPES-KOH pH 7.9, 10 mM KCl, 1 mM CaCl2, 1 mM MnCl2) and bound to the sample for 10 min. Samples were then split into 0.5-ml Lobind tubes, one for each antibody, and resuspended in 25 µl of antibody buffer containing primary antibody at a 1:10 dilution. Samples were incubated at 4 °C overnight, washed twice with 100 µl of wash buffer and then resuspended in 50 µl of wash buffer containing secondary antibody at a 1:50 dilution. Samples were incubated at 4 °C for 1 h, washed twice with 100 µl of wash buffer and then resuspended in 50 µl of 300-wash buffer with a 1:50 dilution of pA-Tn5. Samples were incubated at 4 °C for 1 h, washed twice with 100 µl of 300-wash buffer and then resuspended in 50 µl of tagmentation solution (300-wash buffer with 10 mM MgCl2). Samples were incubated at 37 °C in a thermocycler with a heated lid for 1 h to allow the tagmentation reaction to go to completion. Samples were then washed with 10 mM TAPS to remove any residual salt and resuspended in 10 mM TAPS pH 8.5 containing 1× DAPI and 1× secondary diluent reagent (Takara, 640196) at a concentration of 400 nuclei per µl. Eighty microliters of cell suspension was loaded into 8 wells of a 384-well plate, together with 25 µl of fiducial reagent (Takara, 640196), according to the manufacturer's instructions. Sample suspension (35 nl) was dispensed on the ICELL8 into the nanowells of a 350v Chip (Takara, 640019). The 350v chip was dried and sealed, and cells were centrifuged at 1,200g for 3 min. The chip was then imaged to identify wells containing a single nucleus, and a filter file was prepared. During image processing, 35 nl of 0.19% SDS in TAPS was added to all nanowells on the ICELL8 using an unfiltered dispense. The chip was then dried, sealed and centrifuged at 1,200g for 3 min and heated at 58 °C in a thermocycler with a heated lid for 1 h to release pA-Tn5 and prepare the tagmented chromatin for PCR. Before opening, the chip was centrifuged at 1,200g and 35 nl of 2.5% Triton X-100 neutralization solution was added to all wells containing a single nucleus via a filtered dispense on the ICELL8. The chip was then dried and 35 nl of i5 indices was added via a filtered dispense. The chip was dried and 35 nl of i7 indices was added via a filtered dispense. The chip was dried, sealed and centrifuged at 1,200g for 3 min. Then, 100 nl of KAPA PCR mix (2.775× HiFi buffer, 0.85 mM dNTPs, 0.05 U KAPA HiFi polymerase per µl) (Roche, 07958846001) was added to all wells containing a single nucleus via two 50-nl filtered dispenses. The chip was centrifuged at 1,200g for 3 min, sealed and placed in a thermocycler for PCR amplification using the following conditions: 1 cycle at 58 °C for 5 min; 1 cycle at 72 °C for 10 min; 1 cycle at 98 °C for 45 s; 15 cycles at 98 °C for 15 s, 60 °C for 15 s and 72 °C for 10 s; and 1 cycle at 72 °C for 2 min. The chip was then centrifuged at 1,200g for 3 min into a collection tube (Takara, 640048). To remove residual PCR primers and detergent, the sample was cleaned using two rounds of SPRIselect bead cleanup at a 1.3 to 1 (vol/vol) ratio of beads to sample. Samples were resuspended in 30 µl of 10 mM Tris-HCl pH 8.0, quantified on a TapeStation Bioanalyzer instrument and pooled with bulk samples for sequencing.

DNA sequencing and data processing

The size distribution and molar concentration of libraries were determined using an Agilent 4200 TapeStation. Up to 48 barcoded CUT&RUN libraries or 96 barcoded CUT&Tag libraries were pooled at approximately equimolar concentration for sequencing. Paired-end 2 × 25 bp sequencing on the Illumina HiSeq 2500 platform was performed by the Fred Hutchinson Cancer Research Center Genomics Shared Resources. This yielded 5–10 million reads per antibody. Single-cell CUT&Tag libraries were prepared using unique i5 and i7 barcodes and pooled with bulk samples for sequencing. For 500–1,000 cells, 20 million reads was sufficient to obtain an average of approximately 80% saturation of the estimated library size for each single cell. Paired-end reads were aligned using Bowtie2 version 2.3.4.3 to UCSC hg19 with the following options: --end-to-end --very-sensitive --no-mixed --no-discordant -q --phred33 -I 10 -X 700. Peaks were called using SEACR version 1.3 after combining replicates. We used custom scripts (https://github.com/mpmeers/JanssensEtAl_MPAL) to merge bulk histone modification-specific peak sets, map fragments to merged peak sets and generate PCA and t-SNE plots. All PCA was implemented using the prcomp() function in R version 4.0.0 (https://www.r-project.org/). t-SNE was implemented using the Rtsne() function in Rtsne library version 0.15. We used all principal components explaining greater than 1% of variance as input to Rtsne, and perplexity was set to the nearest integer to the square root of the number of rows in the input matrix. Bivalent gene classifications (H3K4me3 specific, H3K27me3 specific and bivalent) for each cell type were determined by quantifying the number of reads mapping in a 2-kb window around the TSS for every gene and using a two-component Gaussian mixture model as implemented using the normalmixEM() function from mixtools library version 1.2.0 in R to distinguish ‘enriched’ and ‘non-enriched’ sets of genes for each histone mark. Bivalent genes were designated as residing in the enriched Gaussian component for both H3K4me3 and H3K27me3 in the cell type in question.

Identifying KMT2Ar oncoprotein targets

To identify unique KMT2Ar targets, we first generated merged sets of SEACR peaks originating from either N-terminal or C-terminal KMT2A antibody-targeted CUT&RUN in each cell type assayed. We quantified the number of fragments mapping to each peak i from each dataset j and summed reads mapped from the two antibodies targeting the same KMT2A terminus in the same dataset to yield N-terminal (nij) and C-terminal (cij) fragments mapped in each peak, existing in cell type sets Nj and Cj, respectively. We calculated the cell-type-specific ‘N over C ratio’ (NCR) for each peak as follows:

$$NCR_{ij} = \log10((((nij)+\min(Nj))/((cij)+\min(Cj))) \times ECDF((Nj+Cj)/2))((nij+cij)/2)) \quad (1)$$

where min(x) is the minimum value of x across the peak set and ECDF(y)(x) is the empirical cumulative distribution function of set y evaluated at x, as implemented in R version 4.0.0 using the ecdf() function. As illustrated in equation (1), ECDF was used to shrink NCR values toward zero in inverse proportion to

the mean nij + cij signal observed in the peak. KMT2Ar identity was evaluated by fitting a two-component Gaussian mixture model to all NCRj and asserting as true any NCRij that was greater than the mean NCR value of NCRj at which the two fitted Gaussian distributions intersected. As a second filter, the above Gaussian mixture modeling approach was repeated using peak length as an input, and peakij was considered to be a KMT2Ar oncoprotein-specific target only when both NCR and peak length met the cutoffs described above. Gaussian mixture modeling was implemented in R using the normalMixEM() function from mixtools library version 1.2.0. For all peaks assigned as KMT2Ar in any cell type, NCR scores were hierarchically clustered using the hclust() function in R on a Euclidean distance matrix generated by the dist() function.

t-SNE embedding of active and repressed chromatin regions

For histone modification data, peaks were called from merged replicate datasets using SEACR34 version 1.3, and peak sets were merged for each modification across all cell types. We generated matrices of raw read counts mapping in each cell type (columns) to merged peaks (rows) for each modification, and we filtered out instances where counts were lower than any count value whose evaluated ECDF was more than 5% diverged from the predicted ECDF value based on a lognormal fit of the data distribution, using the fitdistr() function from MASS library version 7.3-53 with densfun set to lognormal. We then log10 transformed the results and rescaled columns to z scores. PCA was performed on the resulting transformed matrices using the prcomp() function in R. For t-SNE analysis, all principal components contributing greater than 1% of variance were used as input to the Rtsne() function from Rtsne library version 0.15, with perplexity set as the nearest integer to the square root of the number of peaks and check_duplicates set as false. We used the resulting two-dimensional t-SNE values as input to the densityClust() function from densityClust library version 0.3 and used that output in the findClusters() function, with rho and delta values set to the 95th percentile of all rho and delta values output from densityClust(), respectively. To generate cluster-average heatmaps, scaled count values were averaged by cluster and the resulting matrix was used as input to the heatmap.2() function from gplots library version 3.1.1. PCA and t-SNE plots were generated using ggplot2 library version 3.3.5 (https://ggplot2.tidyverse.org/).

UMAP embedding of single cells

Single cells that did not meet a minimum number of reads (n = 300) or fell below the normal distribution of FRiP values defined by aggregate data were removed. Then, a single-cell count matrix of N features, defined by 5-kb windows tiled across the genome, by M cells was generated. These matrices were binarized and normalized via latent semantic indexing (LSI) (Granja, 2021). The normalized count matrix was reduced from N dimensions to two dimensions using UMAP and plotted. We generated imputed gene scores using MAGIC41 for subsequent analysis. Normalized dispersion was calculated from these gene scores using SCANPY55 version 1.6.0.

Statistical analysis

All comparisons of the normalized AutoCUT&RUN or AutoCUT&Tag signal across peak sets as well as comparisons of normalized dispersion between gene groups were done using two-sample t tests (two sided) with the SciPy.stats.ttest_ind() function in Python; P values were not corrected for multiple-hypothesis testing. Comparisons between the distributions of wide KMT2A peaks and KMT2A oncoprotein-binding sites across gene annotations were done using Fisher's exact tests; P values were similarly not corrected for multiple-hypothesis testing. H3K4me3 peaks that showed a significant change in H3K4me3 signal in response to treatment with 30 μM of the Menin binding inhibitor VTP50469 were identified by DESeq2 version 1.32.0 using the Wald test. Here P values were corrected for multiple-hypothesis testing (adjusted P value) in a manner that was proportional to the number of peaks per sample.

Preparation of figure panels

All heatmaps were generated using DeepTools56 version 3.5.0. t-SNE plots colored by maximum signal from immunophenotype class were generated using ggplot2 version 3.3.5. All data were analyzed using bash, Python (https://github.com/python) or R version 4.0.0. The following packages were used in Python: Matplotlib version 3.2.2, NumPy version 1.18.5, Pandas version 1.0.5, Scipy version 1.5.0, Scanpy version 1.6.0 and Seaborn version 0.10.1.

Data Availability

All primary sequencing data have been deposited as paired-end fastq files in the Gene Expression Omnibus under accession code GSE159608.

Author Contribution

D.H.J. and S.H. optimized the CUT&Tag method for automation, and D.H.J. adapted these modifications for single-cell CUT&Tag profiling. S.M. provided clinical samples and helpful discussion. D.H.J., J.F.S., K.A. and S.H. designed experiments. D.H.J., E.B. and J.F.S. performed experiments. D.H.J., M.P.M., S.J.W. and J.F.S. performed data analysis. D.H.J., M.P.M., K.A. and S.H. wrote the manuscript. All authors read and approved the final manuscript.

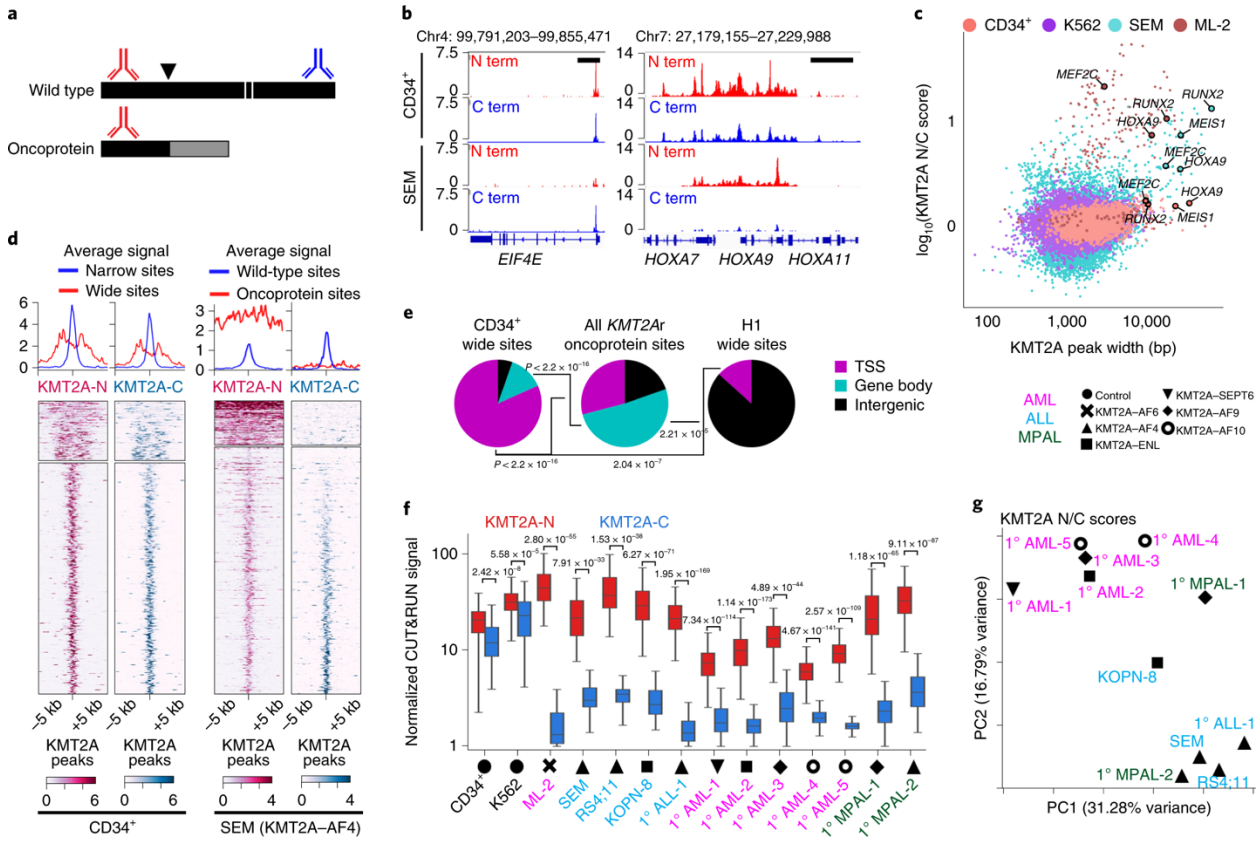


Figure 5.1: AutoCUT&RUN profiling of KMT2A fusion protein binding.
a) A general strategy for mapping KMT2A fusion proteins. The wild-type KMT2A protein (black) is cleaved (white lines) into KMT2A-N and KMT2A-C proteins. Common oncogenic lesions (black arrowhead) produce in-frame translation of oncogenic KMT2A with numerous fusion partners (gray). Antibodies to C-terminal KMT2A (blue) recognize wild-type KMT2A-C. Antibodies to N-terminal KMT2A (red) recognize wild-type KMT2A-N and oncogenic KMT2A fusion proteins. b) Example of a wild-type KMT2A-binding site (EIF4E) and an oncoprotein-binding site (HOXA locus). Black scale bars, 10 kb. c) Scatterplot comparing KMT2A peak width and relative enrichment of the KMT2A N versus C terminus in control (CD34+ and K562) samples and KMT2Ar (SEM and ML-2) samples. d) Heatmap comparison of KMT2A signal over 10-kb windows centered on wide KMT2A-binding sites (top) and narrow KMT2A-binding sites (bottom) in CD34+ HSPCs and KMT2A fusion protein-binding sites (top) and wild-type KMT2A-binding sites (bottom) in SEM cells. e) Pie charts showing the fraction of wide KMT2A peaks (CD34+ and H1 cells) and KMT2A fusion-bound sites (KMT2Ar samples) overlapping transcriptional start sites (TSSs), gene bodies and intergenic regions. P values were computed using Fisher's exact test. f) Box plots of KMT2A-N and KMT2A-C signal showing that the antibody to N-terminal KMT2A is enriched relative to the antibody to the C-terminal portion of KMT2A at fusion-binding sites. The center line indicates the median, box limits represent the first and third quartiles, and whiskers show all data within 1.5 times the interquartile range (IQR) of the lower and upper quartiles; outliers are not shown. P values were computed using a two-sample t test (two sided). CD34+, n = 131; K562, n = 65; ML-2, n = 144; SEM, n = 91; RS4;11, n = 92; KOPN-8, n = 192; 1° ALL-1, n = 156; 1° AML-1, n = 349; 1° AML-2, n = 423; 1° AML-3, n = 103; 1° AML-4, n = 270; 1° AML-5, n = 186; 1° MPAL-1, n = 248; 1° MPAL-2, n = 189. g) Principal-component analysis (PCA) of fusion oncoprotein-binding sites in KMT2Ar samples. The first two components are shown.

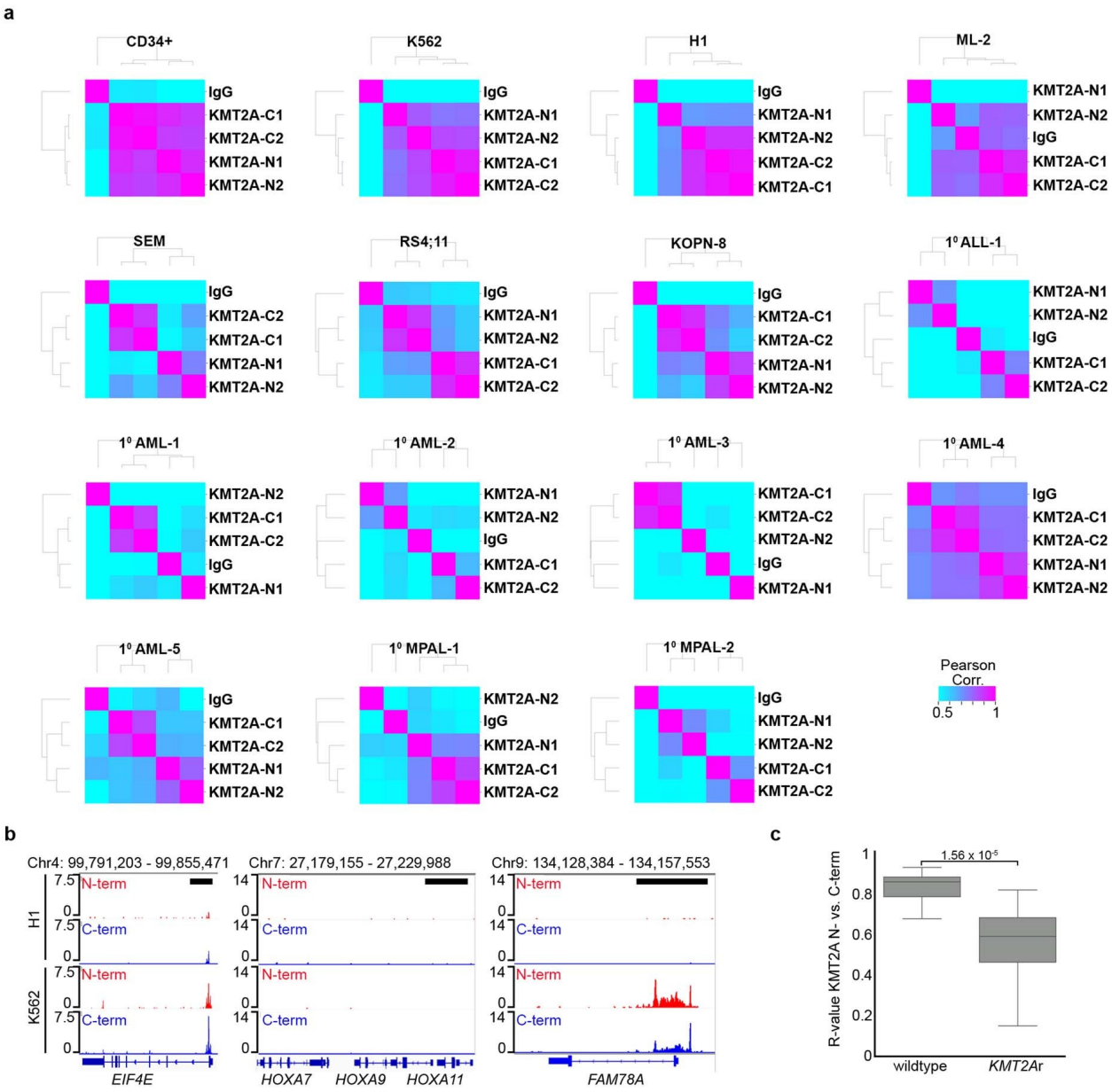


Figure 5.2: KMT2A N-terminus and C-terminus specific antibodies for AutoCUT&RUN chromatin profiling. a) Pearson correlation matrices between KMT2A N-terminus and C-terminus specific antibodies over the merged KMT2A peaks for each sample. In the control CD34 + progenitors, as well as K562 and H1 cells signals for the KMT2A-N1 antibody (Millipore Cat# 05-764), KMT2A-N2 antibody (Cell Signaling Tech Cat# 14689 S), KMT2A-C1 antibody (Millipore Cat# 05-765) and KMT2A-C2 antibody (Santa Cruz Cat# sc-374392) are all highly correlated, indicating that the N-terminal and C-terminal regions of wild-type KMT2A co-localize on chromatin. In the KMT2Ar sample profiles the N-terminal antibodies show higher correlations with one another than they do with the C-terminal antibodies, indicating the KMT2A fusion binding is detected by the N-terminal antibodies and uncoupled from the KMT2A wild-type protein mapped by the C-terminal antibodies. b) Genome browser tracks showing wild-type KMT2A enrichment over the TSS of the EIF4E gene in both H1 and K562 cells. In contrast to CD34 + HSPCs, many critical hematopoietic cell fate determinants (for example HOXA9) are not bound by wild-type KMT2A in H1 or

K562 cells. A broad distribution of KMT2A signal is found across the gene bodies of a limited collection of target genes (for example FAM78A) in K562 cells. Black scale bars = 10 kb. c) The Pearson correlation of KMT2A N- versus C-terminal profiles is significantly higher in the control KMT2A wild-type samples than in the KMT2Ar samples. Center line = median; box limits = first and third quartiles; whiskers show all data within 1.5 IQRs of the lower and upper quartiles; outliers are not shown; P value was computed using a two sample t-test (two-sided); wildtype, n = 12; KMT2Ar, n = 48.

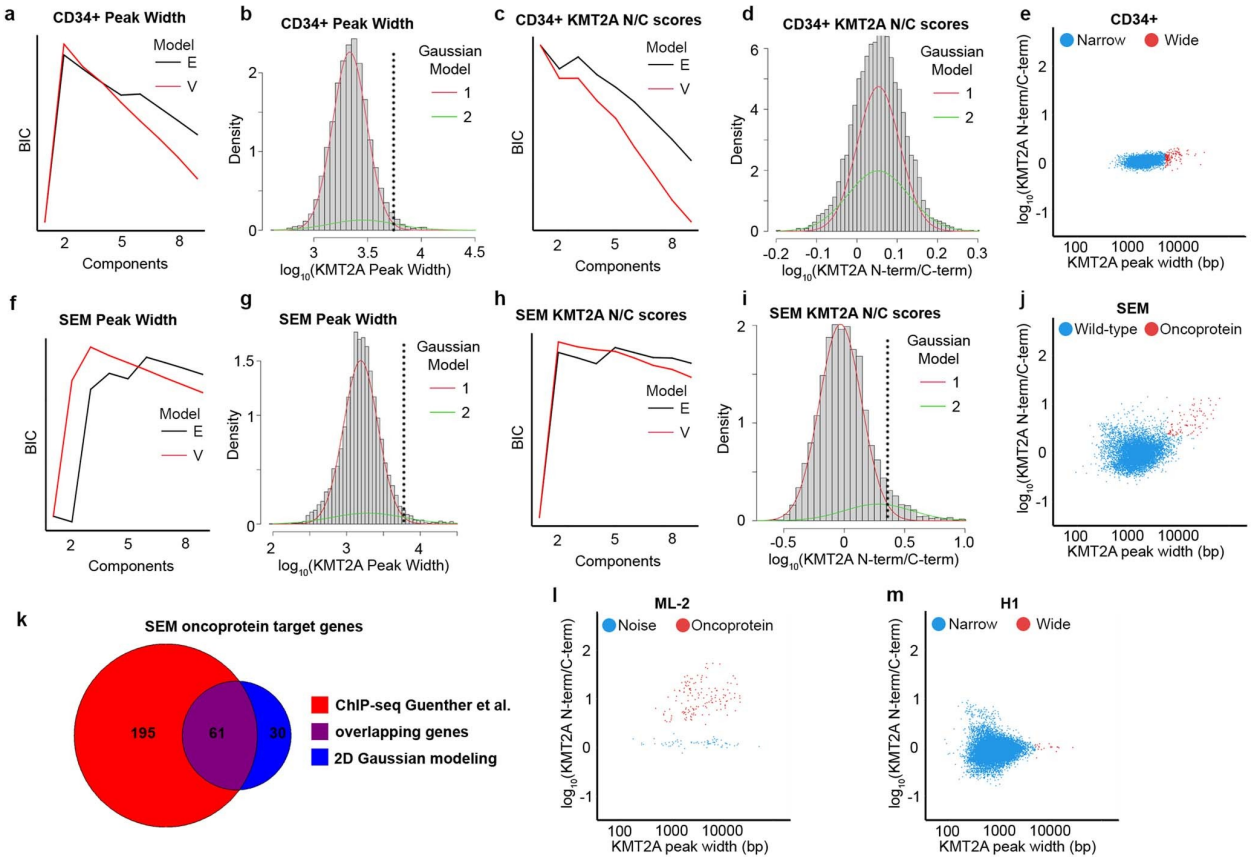


Figure 5.3: A uniform approach to identify sites bound by the KMT2A fusion protein using KMT2A N- and C-terminal profiles.

a) Line plot comparing the Bayesian Information Criterion (BIC) for a range of Gaussian mixture models containing 1-9 components using either equal (E) or unequal (V) variance to model the distribution of KMT2A peak widths in CD34 + HSPCs. A two component Gaussian mixture model provides the highest BIC. b) Histogram of KMT2A peak widths in CD34 + HSPCs showing the two Gaussian models fit to the data. The dotted line indicates the threshold separating peaks called as 'wide' versus 'narrow'. c) Same as (a) but modeling the distribution of the relative enrichment of the KMT2A N- versus C-terminus (KMT2A N/C scores) over KMT2A peaks in CD34 + HSPCs. The highest BIC is achieved by a single Gaussian distribution. d) A two component model fails to partition the KMT2A peaks by N/C scores in CD34 + HSPCs. e) Scatter plots comparing the KMT2A peak width and N/C scores in control CD34 + HSPCs; wide peaks indicated in red. f) Same as (a) but for the KMT2Ar SEM cell line. g) Same as (b) but for SEM cells. h) Same as (c) but for SEM cells. Here, a two component Gaussian mixture model achieves the highest BIC. i) Two Gaussian models fit to the SEM KMT2A N/C scores. The dotted line indicates the threshold separating peaks with 'high' versus 'low' KMT2A N/C scores. j) Same as (e) but for SEM cells; oncoprotein targets are indicated in red. k) Venn diagram of KMT2A-AF4 oncoprotein target genes in SEM cells called using either this two-dimensional Gaussian modeling approach or using ChIP-seq23. l) Same as (e) for the KMT2Ar ML-2 cell line which lacks the wild-type KMT2A allele. m) Same as (e) for the control H1 sample.

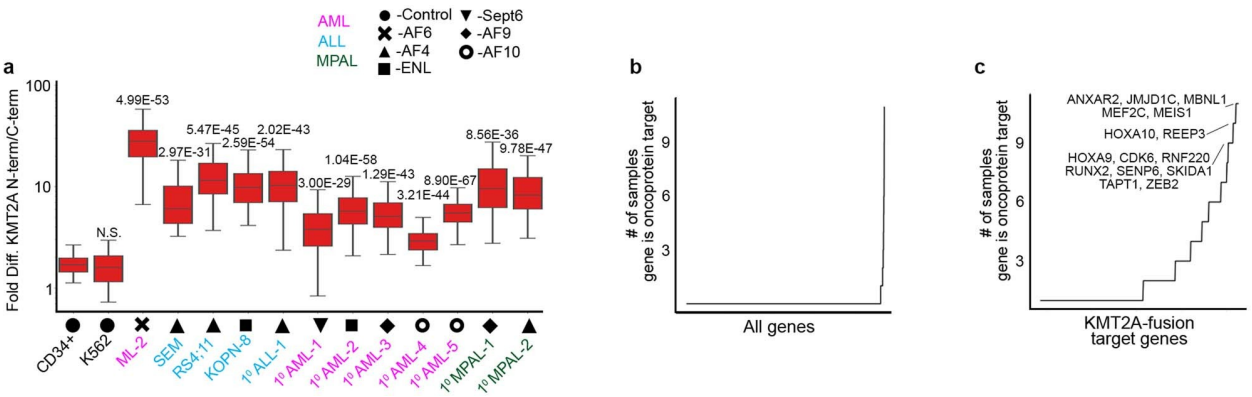


Figure 5.4: Comparison of fusion oncoprotein binding sites across all samples.
a) Box plot showing the fold difference between the KMT2A-N and -C signal at wide KMT2A peaks in the control CD34 + and K562 samples is significantly less than at oncoprotein target sites in the KMT2Ar samples. Center line = median; box limits = first and third quartiles; whiskers show all data within 1.5 IQRs of the lower and upper quartiles; outliers are not shown; P values were computed using a two sample t-test (two-sided) comparing the fold difference between the KMT2A-N and -C signal on KMT2A fusion oncoprotein sites to wide KMT2A peaks in the CD34 + control; CD34 + , n = 131; K562, n = 65; ML-2, n = 144; SEM, n = 91; RS4;11, n = 92; KOPN-8, n = 192; 10 ALL-1, n = 156; 10 AML-1, n = 349; 10 AML-2, n = 423; 10 AML-3, n = 103; 10 AML-4, n = 270; 10 AML-5, n = 186; 10 MPAL-1, n = 248; 10 MPAL-2, n = 189. b) Line plot of all genes showing the number of samples in which a particular gene is called as an oncoprotein target. c) Same as (b) but only genes called as an oncoprotein target in at least one sample are included. The most frequent oncoprotein target genes are indicated.

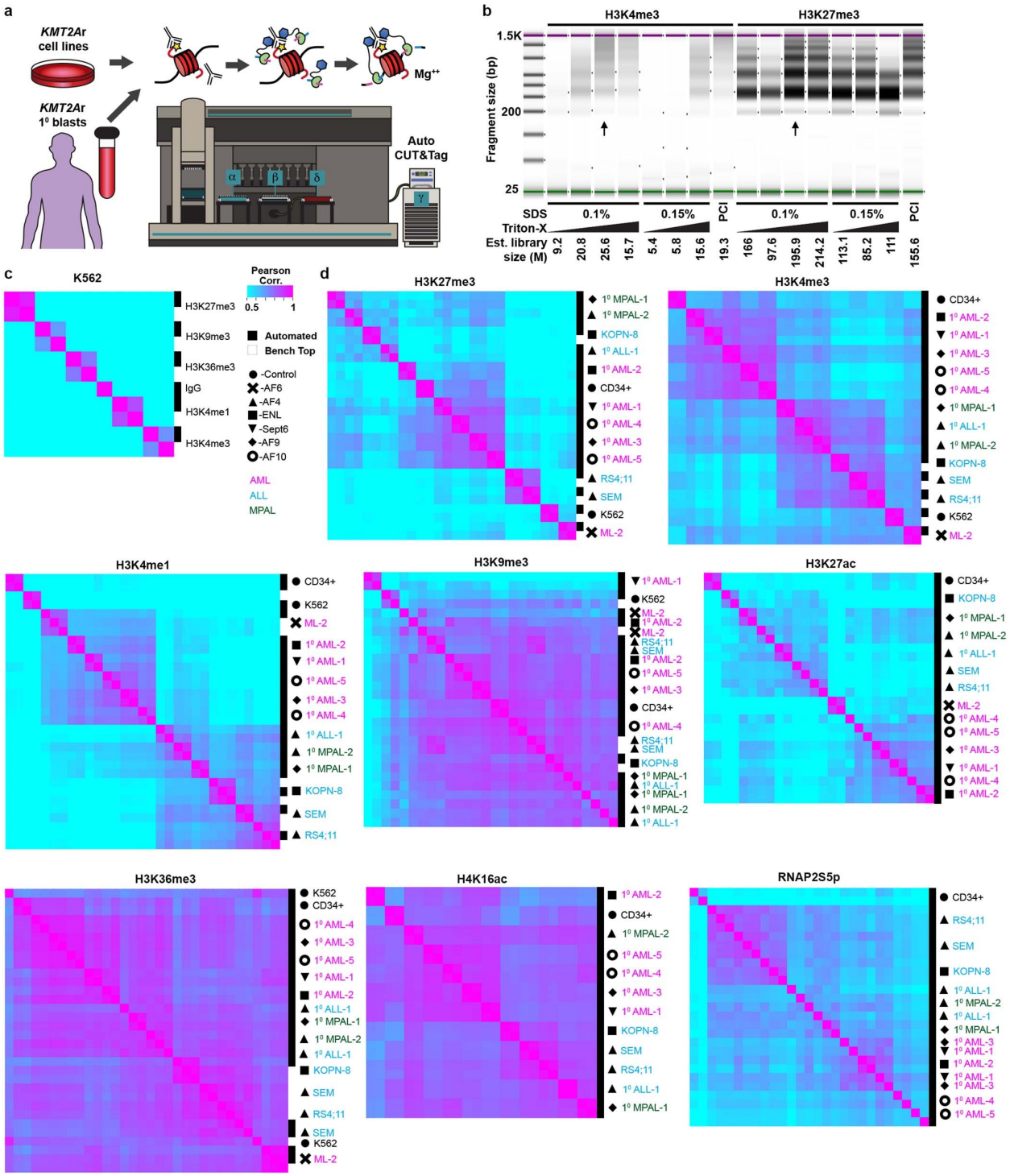


Figure 5.5: Adaptation of CUT&Tag for full automation.
a) Con-A bound nuclei are incubated with the primary antibody of interest and arrayed for AutoCUT&Tag profiling on a liquid handling robot equipped for high volume magnetic separation (α), low volume magnetic separation (β) and temperature control (δ, γ). This method prepares up to 96 sequencing-ready samples in a single day that can all be pooled on a single HiSeq two-lane or comparable flow cell for sequencing. The automated protocol uses a low concentration of SDS to displace bound Tn5 from

tagmented DNA, and Triton-X100 to quench the detergent for PCR. b) CUT&Tag libraries imaged on an Agilent Tapestation 4200 instrument, this High sensitivity D1000 screen tape image is uncropped. To optimize the DNA release and quenching conditions for AutoCUT&Tag varying amounts of SDS and Triton-X100 were tested for library yield. Arrows indicate the optimum condition. c) Pearson correlation matrix of reproducibility between benchtop and automated CUT&Tag profiling methods on K562 fixed nuclei with 5 antibodies to histone modifications as well as the IgG negative control antibody. The log-transformed signal was compared across 5-kb bins tiling the entire genome. d) Pearson correlation matrix of reproducibility between benchtop and automated CUT&Tag profiling methods on fixed nuclei from all 12 KMT2Ar leukemias as well as the CD34 + progenitor and K562 control samples using antibodies against H3K27me3, H3K4me3, H3K4me1, H3K36me3, H3K9me3, H3K27ac, H4K16ac, and RNAP2S5p. Here, log-transformed signal was compared across the merged peak file of all samples for each mark. H3K27me3, H3K4me3, H3K4me1, H3K27ac, and RNAP2S5p show the highest variation between samples, and most reliably cluster sample replicates together.

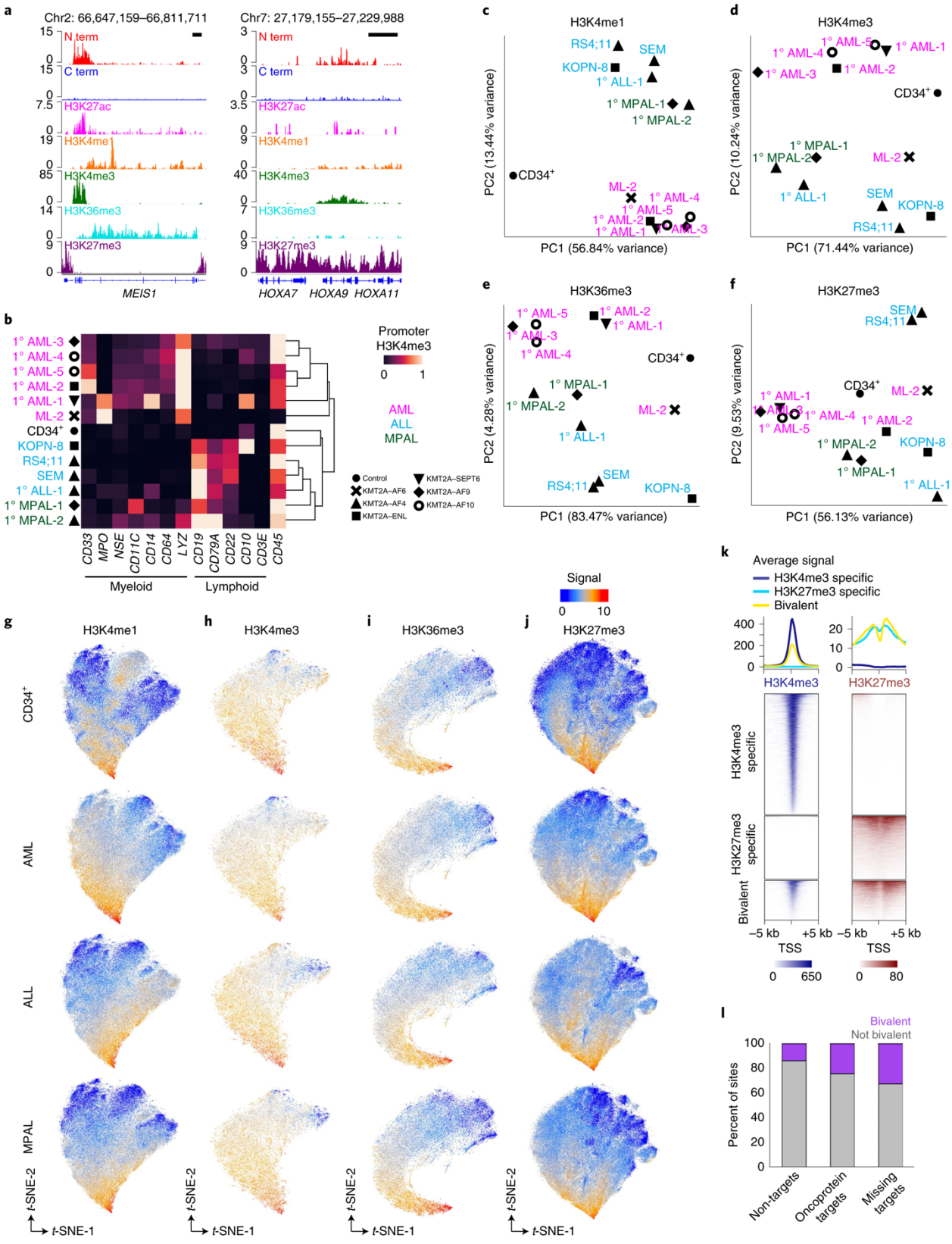


Figure 5.6: Clustering regulatory features distinguishes common and restricted elements in leukemia samples.

a) The MEIS1 locus is a direct target of KMT2A–AF9 in the 1° MPAL-1 sample and is decorated by both active and repressive chromatin marks. The HOXA cluster is relatively repressed in this tumor. Black scale bars, 10 kb. b) H3K4me3 signal at the promoters of diagnostic immunophenotypic markers accurately classifies AML, ALL and MPAL leukemias. c) PCA clustering analysis of H3K4me1-marked regions across the genome separates samples according to lineage specificity. d) Same as in c for H3K4me3. e) Same as in c for H3K36me3. f) Same as in c for H3K27me3. Grouping samples by PCA of H3K27me3-marked repressive chromatin separates tumors of the same lineage. g) Two-dimensional t-distributed stochastic neighbor embedding (t-SNE) projections of all H3K4me1-marked regions separates lineage-specific regulatory elements. Each colored pixel corresponds to a single H3K4me1 peak, colored by the maximum intensity in the indicated sample type. h) t-SNE projection of H3K4me3 identifies lineage-specific promoters. i) Same as in g for H3K36me3. j) Same as in g for H3K27me3. AML and ALL samples display widespread H3K27me3, whereas H3K27me3 is more confined in the CD34+ control and MPAL samples. k) Heatmaps showing H3K4me3- and H3K27me3-specific regions (top) as well as regions called as bivalent (bottom) in the 1° MPAL-1 sample. l) Comparison of bivalent chromatin at genes not bound by the KMT2A fusion (non-targets), genes bound by the KMT2A fusion (oncoprotein targets) and genes bound in the majority of samples but missed in the sample of interest (missing targets) shows that a bivalent chromatin signature is enriched at oncoprotein target genes.

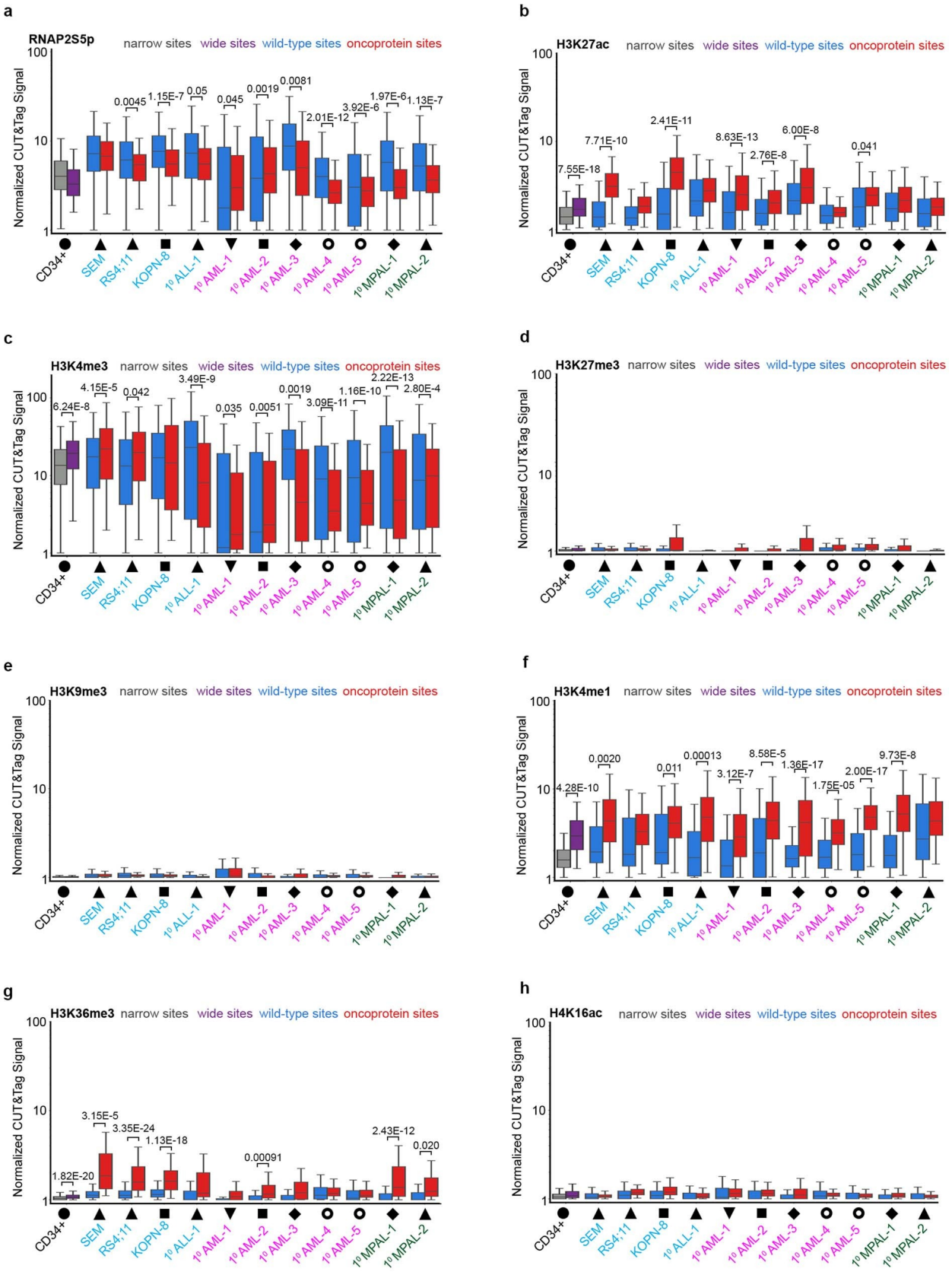


Figure 5.7: Chromatin features of KMT2A fusion protein binding sites.

a) Quantification of the RNAP2Sp signal across the KMT2A wild-type and oncoprotein target sites in all samples. Signal is normalized by total coverage, as well as peak width. For all box plots center line = median; box limits = first and third quartiles; whiskers show all data within 1.5 IQRs of the lower and upper quartiles; outliers are not shown; P values were computed using a two sample t-test (two-sided). N values listed in order: narrow sites, wide sites for the control, or wild-type sites, oncoprotein sites for the KMT2Ar samples. CD34 + , n = 6204, 131; SEM, n = 8168, 91; RS4;11, n = 10287, 92; KOPN-8, n = 9723, 192; 10 ALL-1, n = 3747, 156; 10 AML-1, n = 15915, 349; 10 AML-2, n = 26148, 423; 10 AML-3, n = 3943, 103; 10 AML-4, n = 7218, 270; 10 AML-5, n = 11256, 186; 10 MPAL-1, n = 8641, 248; 10 MPAL-2, n = 15179, 189. b) Same as (a) but for H3K27ac. H3K27ac is significantly enriched at the fusion oncoprotein binding sites in the SEM, KOPN-8, 10 AML-1, 10 AML-2, 10 AML-3 and 10 AML-5 KMT2Ar leukemia samples as compared to the sample matched wild-type KMT2A bound sites. c) Same as (a) but for H3K4me3. H3K4me3 is significantly enriched at the fusion oncoprotein binding sites in SEM, RS4;11, 10 AML-1, 10 AML-2 and 10 MAPL-2 and significantly depleted in 10 ALL-1, 10 AML-3, 10 AML-4, 10 AML-5 and 10 MPAL-1. d) Same as (a) but for H3K27me3. e) Same as (a) but for H3K9me3. f) Same as (a) but for H3K4me1. g) Same as (a) but for H3K36me3. h, Same as (a) but for H4K16ac.

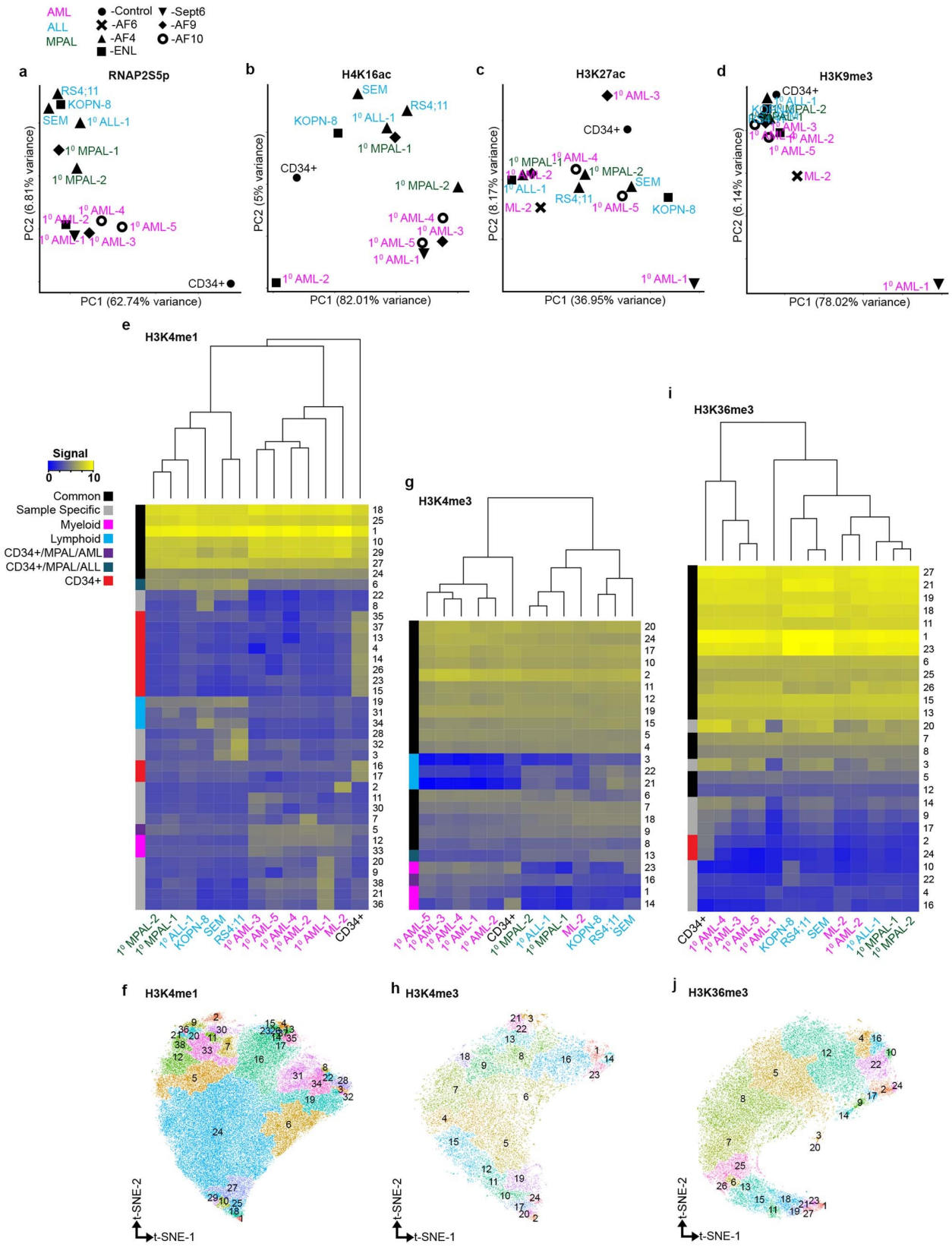


Figure 5.8: Grouping KMT2Ar samples and regulatory regions according to the sample specific AutoCUT&Tag chromatin profiles.

a) PCA of genome-wide RNAP2S5p signal separates samples by leukemia subtype. b) PCA of genome-wide H4K16ac signal separates samples by leukemia subtype. c) Samples organized by PCA of H3K27ac are not separated by leukemia subtype. d) PCA of genome-wide H3K9me3 signal only distinguishes the 10 AML-1 KMT2A-Sept6 containing sample from the rest. e) Clustering analysis separates the H3K4me1-marked regions into 38 groups, and the heatmap shows the average signal intensity of H3K4me1 in each of these groups (y-axis) for each KMT2Ar leukemia sample (x-axis). The colors alongside the heatmap show the subtype designation of each group as common across samples (black), myeloid (magenta), lymphoid (cyan), shared by CD34 + , MPAL and myeloid samples (plum), shared by CD34 + , MPAL and lymphoid samples (teal), CD34 + specific (red), or sample specific (gray). f) Two-dimensional t-SNE projections of all H3K4me1-marked regions as in Figure 5.6g, but colored according to the group designation as indicated by the numbers along the right side of (e). g) Same as (e) for H3K4me3. Clustering analysis separates the H3K4me3-marked regions into 24 groups. h) Same as (f) for H3K4me3. i) Same as (e) for H3K36me3. Clustering analysis separates the H3K36me3-marked regions into 27 groups. j) Same as (f) for H3K36me3.

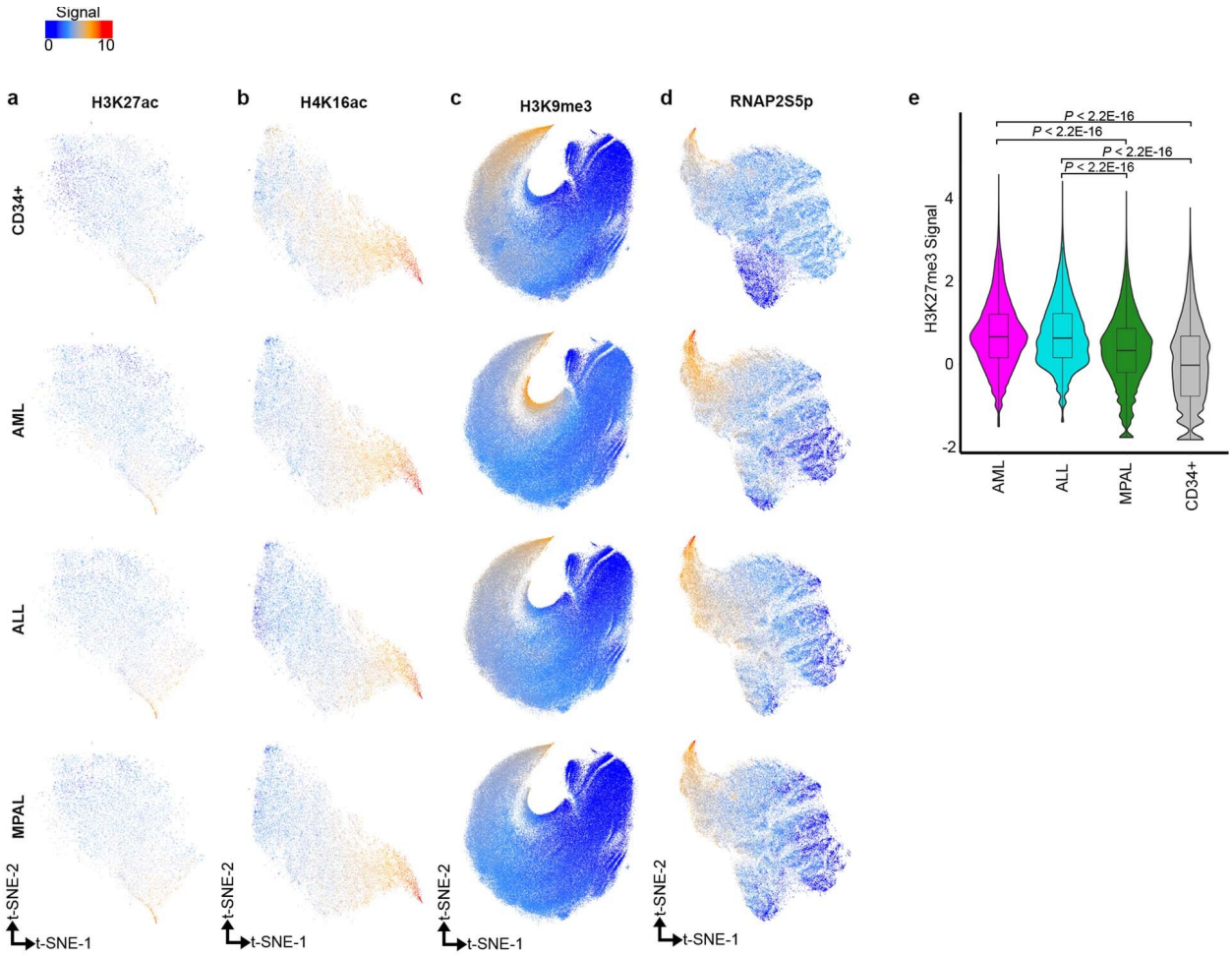


Figure 5.9: The genome wide distributions of additional active and repressive modifications profiled by AutoCUT&Tag.
a-c) t-SNE projections of H3K27ac, H4K16ac and H3K9me3 samples do not organize the enriched elements based on lineage specificity. d) t-SNE projection of the RNAP2S5p bound regions identifies groups of elements that show CD34 + , AML, ALL and MPAL specific enrichment of RNAP2S5p. e) Violin plot comparing the H3K27me3 signal across the composite H3K27me3 peak set identified in all samples. The H3K27me3 signal is significantly lower in the CD34 + and MPAL samples than in the AML and ALL samples, indicating that a greater proportion of the genome is marked by H3K27me3 in the more differentiated AML and ALL samples. Center line = median; box limits = first and third quartiles; whiskers show all data within 1.5 IQRs of the lower and upper quartiles; outliers are not shown; P values were computed using a two sample t-test (two-sided); for all samples n = 89258 H3K27me3 marked regions.

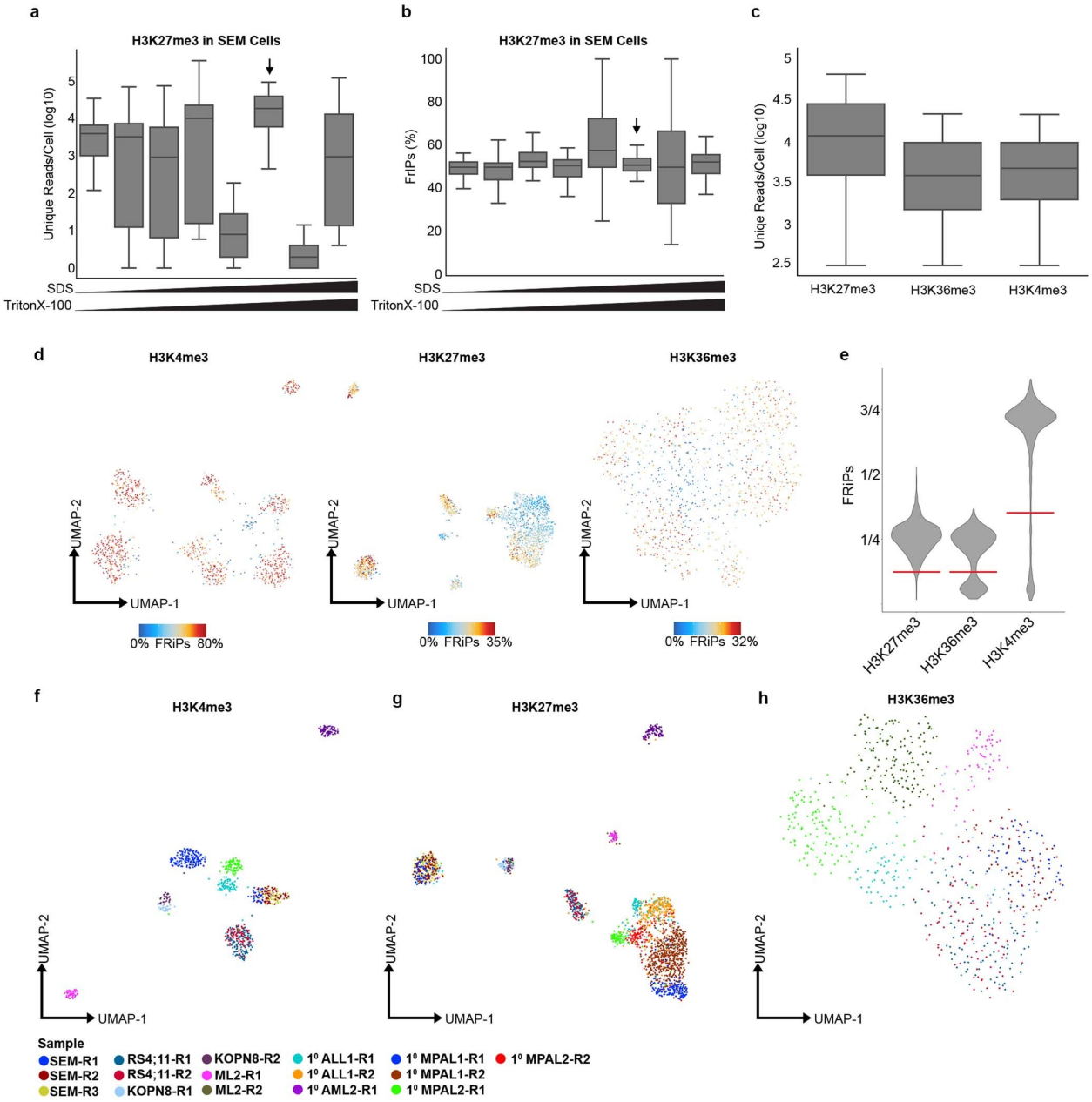


Figure 5.10: Optimization of CUT&Tag-Direct for single cell applications on the ICELL8.
a) Titrating the concentration of SDS and Triton-X in the nanowell increases the library yield for individual cells, and identifies optimum conditions for Tn5 release and PCR enrichment (Arrow). For all box plots center line = median; box limits = first and third quartiles; whiskers show all data within 1.5 IQRs of the lower and upper quartiles; outliers are not shown; P values were computed using a two sample t-test (two-sided); from left to right n = 84, 63, 51, 71, 37, 62, 31, 39 cells. b) The Fraction of Reads in Peaks (FRiPs) varies across SDS and Triton-X titration conditions. Arrow indicates the optimum conditions. N values the same as in (a). c) Boxplot of unique reads per cell across all of the cells profiled for H3K27me3, H3K36me3 and H3K4me3; H3K27me3, n = 3,611 cells; H3K36me3, n = 1,137 cells; H3K4me3, n = 1,528 cells. d) UMAP projection of single-cells profiled with H3K4me3, H3K27me3 or H3K36me3 colored according to the FRiP scores of each individual cell using peaks called on the aggregate data of all cells profiled for a given mark. Cells with low FRiP scores tend to fall in between clusters and were removed as a quality control. e) Violin plot of FRiP scores for all individual cells profiled

using the indicated histone mark; red lines indicate quality control cut-offs for each mark. N values same as (c). f) UMAP projection of single cells profiled for H3K4me3 and colored according to batch (R1, R2, R3). Replicates profiled on different days group together in UMAP space indicating that batch effects have a minimal impact on clustering cells according to the H3K4me3 profiles. g) Same as (f) for H3K27me3. h) Same as (f) for H3K36me3.

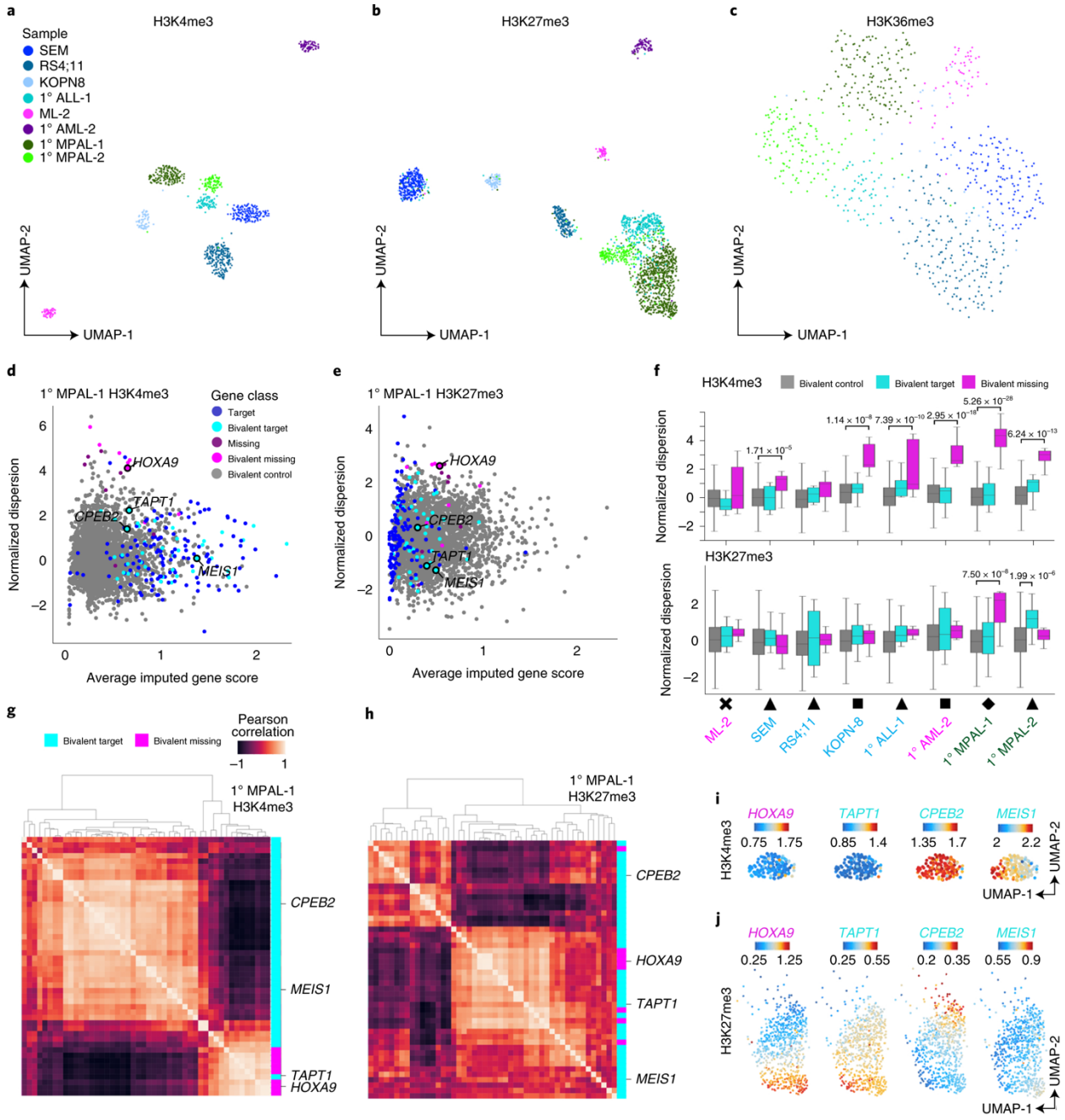


Figure 5.11: Single-cell profiling of H3K4me3 and H3K27me3 reveals chromatin heterogeneity at KMT2A fusion target loci.

a) UMAP projection of the H3K4me3 profiles in single leukemia cells resolves sample-specific clusters. b) Same as in a for H3K27me3. A fraction of cells in the 1° ALL-1, 1° MPAL-1 and 1° MPAL-2 samples intermingle in H3K27me3 UMAP space. c) Same as in a for H3K36me3. Leukemia cells do not form tight sample-specific groups according to H3K36me3 profile. d) Scatterplot comparing the 1° MPAL-1 average imputed H3K4me3 scores and normalized dispersion of genes grouped according to KMT2A fusion-binding status (target, missing target or unbound control) and promoter bivalency status in bulk profiling assays. Select genes are highlighted. e) Same as in d for H3K27me3. f) KMT2A fusion targets show elevated H3K4me3 and H3K27me3 dispersion across select leukemia samples. The center line indicates

the median, box limits represent the first and third quartiles, and whiskers show all data within 1.5 times the IQR of the lower and upper quartiles; outliers are not shown. P values were computed using a two-sample t test (two sided). For each sample, n values are listed for bivalent controls, bivalent targets and bivalent missing targets: ML-2: n = 2,425, 14, 14; SEM: n = 3,441, 8, 15; RS4;11: n = 2,635, 6, 9; KOPN-8: n = 2,044, 31, 5; 1° ALL-1: n = 1,873, 6, 9; 1° AML-2: n = 2,483, 17, 12; 1° MPAL-1: n = 3,101, 40, 8; 1° MPAL-2: n = 3,018, 15, 7. g) Organizing genes according to the covariance of H3K4me3 imputed gene scores across 1° MPAL-1 cells resolves groups that vary in concert with one another from cell to cell but are anticorrelated with genes in the other group. Selected genes are highlighted for comparison. h) Same as in g for H3K27me3. i) Imputed H3K4me3 gene scores of the highlighted genes from g and h displayed on a UMAP plot of 1° MPAL-1 cells shown as a dark green cluster in a. j) Same as in i for H3K27me3. HOXA9 and TAPT1 are representative of one group of covariant KMT2A fusion target genes, which are divergent from the second group (for example, CPEB2 and MEIS1).

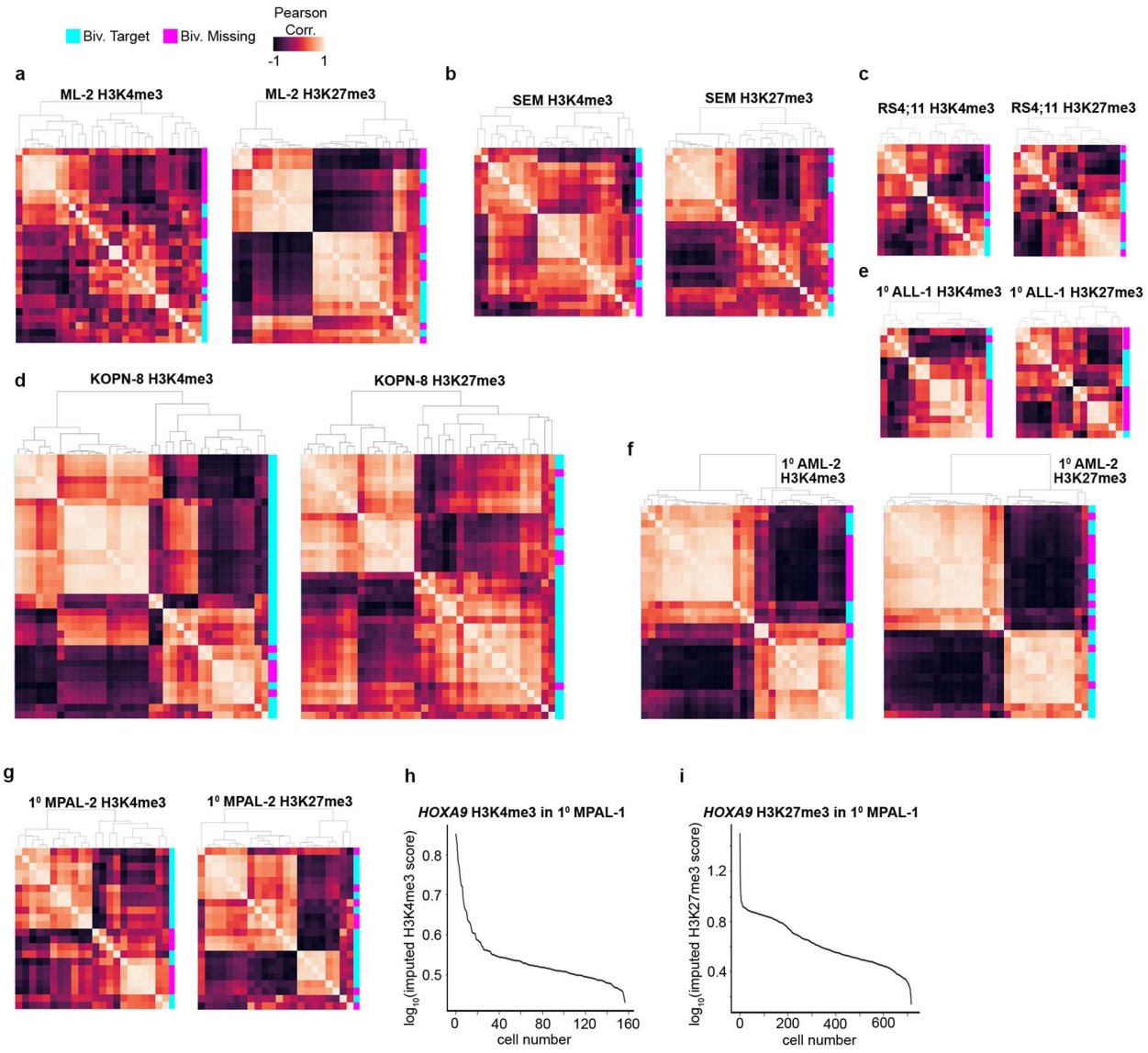


Figure 5.12: Networks of KMT2A fusion target genes show divergent patterns of active and repressive chromatin marks within the same leukemia.
a) Organizing genes according to the co-variance of either the imputed H3K4me3 (left) or H3K27me3 (right) scores across ML-2 cells resolves groups that vary in concert with one another from cell-to-cell but are anti-correlated with genes in the other group. Bivalent targets (Cyan) and bivalent missing targets (Magenta) are indicated. b) Same as (a) for SEM cells. c) Same as (a) for RS4;11 cells. d) Same as (a) for KOPN-8 cells. e) Same as (a) for the 10 ALL-1 cells. f) Same as (a) for the 10 AML-2 cells. g) Same as (a) for the 10 MPAL-2 cells. The ML-2, KOPN-8, 10 AML-2 and 10 MPAL-2 cells show the clearest distinction between divergent groups of oncoprotein target genes. h) Knee plot showing the distribution of the imputed HOXA9 H3K4me3 scores across all cells profiled from the 10 MPAL-1 sample. Only ~15% of cells have a log-transformed HOXA9 H3K4me3 score > 0.6. i) Same as (h) but showing imputed HOXA9 H3K27me3 scores. The majority of 10 MPAL-1 cells (~55%) have a log-transformed HOXA9 H3K27me3 score > 0.6.

Chapter 6

Conclusion

In my doctoral research, I integrated data across the modalities of histone modifications to generate a simple model that resolves cell type specific gene regulation. I then transitioned from an in silico to an experimental approach to uncover true cell-type specific gene regulation without the obscurity of sample heterogeneity by adapting CUT&Tag to single-cell resolution. This study has demonstrated a proof-of-concept for robust profiling of both active and repressive histone marks in single cells. Furthermore, this study revealed that scCUT&Tag is resolute enough to uncover the single cell temporal dynamics of stem cell differentiation.

My thesis work is part of a large-scale effort to build a human cell atlas to better understand human health and improve disease diagnosis and treatment. This study contributed to the human cell atlas project by developing scCUT&Tag, a technology that profiles any protein-DNA interaction in single-cell resolution. Furthermore we used scCUT&Tag to generate thousands of single-cell chromatin profiles for healthy and diseased human tissues to expand the human cell atlas. In the following sections of this chapter, I will review the limitations of scCUT&Tag and how it compares to the newest emerging sequencing technologies in terms of single-cell yield and 'multi-omics'.

Single cell scalability & yield

One limitation of single-cell profiling, although not specific to CUT&Tag, is that the data generated is sparse (Lähnemann et al., 2020). As such, we can only generate data driven hypotheses about what we directly observe. We can also make biological inferences from what's not present in the data but resolving a statistically significant biological event is difficult due to data drop-out rather than an unobserved biological phenomenon. The simplest way to overcome this obstacle is to increase read-out per single cell through modifying the experimental protocol. The less obvious approach to overcome sparsity is to rely on the concept of "meta-cells", where similar cells are aggregated for subsequent analysis (Baran et al., 2019). From this approach, even if individual single-cells have low read-out, sparsity can be reduced by profiling and aggregating more cells.

Prior works were only able to profile hundreds of single-cell DNA-protein profiles, but we demonstrated an order of magnitude increase in single cells per experiment (Hainer et al., 2018; Rotem, 2015). Approximately ~30,000 single-cells were profiled in total for my thesis work across all studies (Janssens, Meers, et al., 2021; Kaya-Okur, 2019; S. J. Wu et al., 2021). However, DNA-protein profiling assays are still orders of magnitude behind scATAC-seq and scRNA-seq. In comparison, a recently

published manuscript profiled millions of single cells for whole organisms with both scATAC-seq and scRNA-seq using combinatorial indexing (Cao et al., 2020; Domcke et al., 2020).

Combinatorial indexing with tagmentation based workflows for DNA-protein profiling has been demonstrated for tens-of-thousands of single cells (Cusanovich et al., 2015). In theory, protein-DNA assays should be scalable to profiling millions of cells, but a major obstacle lies in the complexity of such workflows. CUT&Tag is significantly simpler to perform than ChIP-seq, but since it targets a specific protein interaction using an antibody there are more steps than ATAC-seq and RNA-seq (Kaya-Okur, 2019). Each additional step increases the number of cells lost through the assay and at a certain point in CUT&Tag the cells aggregate together and clump, thus lowering the yield of each experiment. Improving the scalability of scCUT&Tag can be done by either reducing the number of steps in the assay or swapping buffer conditions to stop cell lysis and subsequent clumping (Janssens, Meers, et al., 2021).

While CUT&Tag can theoretically profile any protein-DNA interactions, in my dissertation I only profiled histone modifications in single-cell resolution and not proteins such as transcription factors because they have low expression. When transitioning from a bulk to a single-cell based assay the data becomes sparse (Lähnemann et al., 2020). To circumvent sparsity, we only profiled the most abundant proteins in chromatin such as histone modifications to increase our reads per single cell. There are already publications demonstrating proof-of-concept for profiling transcription factors in single-cell resolution, but the yield is around ~50-100 reads per single-cell (Bartosovic et al., 2021). The question now is whether the low reads per cell in TFs are even useful for identifying cellular subpopulations or understanding new biology. Regardless, the low reads per cell is indicative of the fact that there's still room for improvement in the single-cell epigenomic profiling.

Multi-omic

The single-cell field as a whole is moving toward a multi-omic approach to stave off data sparsity. The goal of multi-omics is to generate more realistic models of gene regulation that capture different modes of transcriptional regulation and processing. The modes of multi-omic data span from transcription to translation, encompassing the central dogma of biology. There are two avenues of approach for multi-omic studies – experimental and in silico.

The first proof-of-concept single-cell multi-omic assay, sci-CAR, integrated single-cell RNA-seq and ATAC-seq simultaneously in a single-cell (Cao et al., 2018). Since then, joint single-cell RNA-seq and ATAC-seq assays have become commercialized by 10x genomics and are becoming the norm. Additionally, in the couple of months since our scCUT&Tag manuscript was published, multiple labs have published on integrating scCUT&Tag with another 'omic' method. To date CUT&Tag has been integrated with spatial technology, scRNA-seq and CITE-seq to name a few (Deng et al., 2021; B. Zhang et al., 2021; Zhu et al., 2021). There have also been efforts to make multiple intra-chromatin measurements in the same single-cell. The most notable being CUT&Tag2for1 and scGET-seq which profiles accessibility and chromatin silencing (Janssens, Otto, et al., 2021; Tedesco et al., 2021). Finally, the newest emerging

iteration of CUT&Tag, dubbed MulTI-Tag, is able to profile up to three protein-DNA interactions in the same cell simultaneously (Gopalan et al., 2021; Meers et al., 2021). Only three interactions have been demonstrated with MulTI-Tag, but the technology can in theory scale up to infinite proteins that are bound to DNA.

An alternative take on multi-omics is an in silco approach where datasets across different assays for a similar population of cells are harmonized. One jarring weakness of this approach is that the actual data measurements aren't performed experimentally in the same cell. Furthermore, different assays yield different cell-type resolution, so matching datasets by cell-types remains quite tricky. Practically, it is easier to perform a single-omic assay and then harmonize the datasets. There are multitudes of in silico approaches for data integration across scRNA-seq and scATAC-seq. However, the use of gene activity scores and weighted nearest neighbor has seen much success (Granja, 2021; Stuart, 2019). Additionally, the use of semi-supervised reference maps learned from CITE-seq data has shown promise to improve the interpretability of scCUT&Tag data (B. Zhang et al., 2021). Ultimately, multi-omics will be an essential tool to study gene regulation and move our understanding of human health and disease forward.

References

10xgenomics. Https://www.10xgenomics.com/solutions/single-cell-atac. (n.d.). https://www.10xgenomics.com/solutions/single-cell-atac

Aldridge, S., Watt, S., Quail, M. A., Rayner, T., Lukk, M., Bimson, M. F., Gaffney, D., & Odom, D. T. (2013). AHT-ChIP-seq: A completely automated robotic protocol for high-throughput chromatin immuno precipitation. *Genome Biol*, *14*. https://doi.org/10.1186/gb-2013-14-11-r124

Alexander, T. B. (2018). The genetic basis and cell of origin of mixed phenotype acute leukaemia. *Nature*, *562*.

Altschuler, S. J., & Wu, L. F. (2010). Cellular heterogeneity: When do differences make a difference? *Cell*, *141*(4), 559–563. https://doi.org/10.1016/j.cell.2010.04.033

Amini, S. (2014). Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nat. Genet.*, *46*. https://doi.org/10.1038/ng.3119

An integrated encyclopedia of DNA elements in the human genome. (2012). *Nature*, *489*. https://doi.org/10.1038/nature11247

Andersson, A. K. (2015). The landscape of somatic mutations in infant MLL-rearranged acute lymphoblastic leukemias. *Nat. Genet.*, *47*.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., & Suzuki, T. (2014). An atlas of active enhancers across human cell types and tissues. *Nature*, *507*. https://doi.org/10.1038/nature12787

Antunes, E. T. B., & Ottersbach, K. (2020). The MLL/SET family and haematopoiesis. *Biochim. Biophys. Acta Gene Regul. Mech.*, *1863*.

Bannister, A. J., & Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell Research*, *21*(3), 381–395. https://doi.org/10.1038/cr.2011.22

Baran, Y., Bercovich, A., Sebe-Pedros, A., Lubling, Y., Giladi, A., Chomsky, E., Meir, Z., Hoichman, M., Lifshitz, A., & Tanay, A. (2019). MetaCell: Analysis of single-cell RNA-seq data using K-nn graph partitions. *Genome Biology*, *20*(1), 206. https://doi.org/10.1186/s13059-019-1812-2

Becht, E. et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nat. Biotechnol.https://doi.org/10.1038/nbt.4314 (2018). (n.d.).

Bernstein, B. E., Mikkelsen, T. S., Xie, X., Kamal, M., Huebert, D. J., Cuff, J., Fry, B., Meissner, A., Wernig, M., & Plath, K. (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, *125*. https://doi.org/10.1016/j.cell.2006.02.041

Bhaduri, A. (2020). Outer radial glia-like cancer stem cells contribute to heterogeneity of glioblastoma. *Cell Stem Cell*, *26*. https://doi.org/10.1016/j.stem.2019.11.015

Bolouri, H. (2018). The molecular landscape of pediatric acute myeloid leukemia reveals recurrent structural alterations and age-specific mutational interactions. *Nat. Med.*, *24*.

Buenrostro, J. D. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, *523*. https://doi.org/10.1038/nature14590

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & 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*. https://doi.org/10.1038/nmeth.2688

Calo, E., & Wysocka, J. (2013). Modification of enhancer chromatin: What, how, and why? *Mol Cell*, *49*. https://doi.org/10.1016/j.molcel.2013.01.038

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

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., & Shendure, J. (2018). Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science (New York, N.Y.)*, *361*(6409), 1380–1385. https://doi.org/10.1126/science.aau0730

Cao, J., O'Day, D. R., Pliner, H. A., Kingsley, P. D., Deng, M., Daza, R. M., Zager, M. A., Aldinger, K. A., Blecher, R., Zhang, F., Spielmann, M., Palis, J., Doherty, D., Steemers, F. J., Glass, I. A., Trapnell, C., & Shendure, J. (2020). A human cell atlas of fetal gene expression. *Science (New York, N.Y.)*, *370*(6518), eaba7721. https://doi.org/10.1126/science.aba7721

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., & Shendure, J. (2019). The single cell transcriptional landscape of mammalian organogenesis. *Nature*, *566*(7745), 496–502. https://doi.org/10.1038/s41586-019-0969-x

Cedar, H., & Bergman, Y. (2009). Linking DNA methylation and histone modification: Patterns and paradigms. *Nature Reviews Genetics*, *10*(5), 295–304. https://doi.org/10.1038/nrg2540

Chan, A. K. N., & Chen, C. W. (2019). Rewiring the epigenetic networks in MLL-rearranged leukemias: Epigenetic dysregulation and pharmacological interventions. *Front. Cell Dev. Biol.*, *7*.

Chu, L.-. F. (2016). Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biol.*, *17*. https://doi.org/10.1186/s13059-016-1033-x

Corces, M. R., Buenrostro, J. D., Wu, B., Greenside, P. G., Chan, S. M., Koenig, J. L., Snyder, M. P., Pritchard, J. K., Kundaje, A., Greenleaf, W. J., Majeti, R., & Chang, H. Y. (2016). Lineage-specific and single cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nature Genetics*, *48*(10), 1193–1203. https://doi.org/10.1038/ng.3646

Corces, M. R., Granja, J. M., Shams, S., Louie, B. H., Seoane, J. A., Zhou, W., Silva, T. C., Groeneveld, C., Wong, C. K., & Cho, S. W. (2018). The chromatin accessibility landscape of primary human cancers. *Science*, *362*. https://doi.org/10.1126/science.aav1898

Core, L. J. (2014). Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.*, *46*. https://doi.org/10.1038/ng.3142

Cotney, J., Muhle, R. A., Sanders, S. J., Liu, L., Willsey, A. J., Niu, W., Liu, W., Klei, L., Lei, J., & Yin, J. (2015). The autism-associated chromatin modifier CHD8 regulates other autism risk genes during human neurodevelopment. *Nat Commun*, *6*. https://doi.org/10.1038/ncomms7404

Couturier, C. P. (2020). Single-cell RNA-seq reveals that glioblastoma recapitulates a normal neurodevelopmental hierarchy. *Nat. Commun.*, *11*. https://doi.org/10.1038/s41467-020-17186-5

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., Lodato, M. A., Frampton, G. M., & Sharp, P. A. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci USA*, *107*. https://doi.org/10.1073/pnas.1016071107

Cui, K. (2009). Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation. *Cell Stem Cell*, *4*.

Cusanovich, D. A., Daza, R., Adey, A., Pliner, H. A., Christiansen, L., Gunderson, K. L., Steemers, F. J., Trapnell, C., & Shendure, J. (2015). Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science (New York, N.Y.)*, *348*(6237), 910–914. https://doi.org/10.1126/science.aab1601

Deng, Y., Zhang, D., Liu, Y., Su, G., Enninful, A., Bai, Z., & Fan, R. (2021). Spatial Epigenome Sequencing at Tissue Scale and Cellular Level. *BioRxiv*, 2021.03.11.434985. https://doi.org/10.1101/2021.03.11.434985

Dijk, D. (2018). Recovering gene interactions from single-cell data using data diffusion. *Cell*, *174*.

Domcke, S., Hill, A. J., Daza, R. M., Cao, J., O'Day, D. R., Pliner, H. A., Aldinger, K. A., Pokholok, D., Zhang, F., Milbank, J. H., Zager, M. A., Glass, I. A., Steemers, F. J., Doherty, D., Trapnell, C., Cusanovich, D. A., & Shendure, J. (2020). A human cell atlas of fetal chromatin accessibility. *Science (New York, N.Y.)*, *370*(6518), eaba7612. https://doi.org/10.1126/science.aba7612

Dou, Y. (2005). Physical association and coordinate function of the H3 K4 methyltransferase MLL1 and the H4 K16 acetyltransferase MOF. *Cell*, *121*.

ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature489, 57–74 (2012). (n.d.).

Erb, M. A. (2017). Transcription control by the ENL YEATS domain in acute leukaemia. *Nature*, *543*.

Ernst, P., Wang, J., Huang, M., Goodman, R. H., & Korsmeyer, S. J. (2001). MLL and CREB bind cooperatively to the nuclear coactivator CREB-binding protein. *Mol. Cell. Biol.*, *21*.

Feinberg, A. P. (2018). The key role of epigenetics in human disease prevention and mitigation. *N Engl J Med*, *378*. https://doi.org/10.1056/NEJMra1402513

Filbin, M. G., Tirosh, I., Hovestadt, V., Shaw, M. L., Escalante, L. E., Mathewson, N. D., Neftel, C., Frank, N., Pelton, K., & Hebert, C. M. (2018). Developmental and oncogenic programs in H3K27M gliomas dissected by single-cell RNA-seq. *Science*, *360*. https://doi.org/10.1126/science.aao4750

Forgione, M. O., McClure, B. J., Eadie, L. N., Yeung, D. T., & White, D. L. (2020). KMT2A rearranged acute lymphoblastic leukaemia: Unravelling the genomic complexity and heterogeneity of this high-risk disease. *Cancer Lett.*, *469*.

Gao, R., Kim, C., Sei, E., Foukakis, T., Crosetto, N., Chan, L.-K., Srinivasan, M., Zhang, H., Meric-Bernstam, F., & Navin, N. (2017). Nanogrid single-nucleus RNA sequencing reveals phenotypic diversity in breast cancer. *Nature Communications*, *8*(1), 228. https://doi.org/10.1038/s41467-017-00244-w

Gardner, R. (2016). Acquisition of a CD19-negative myeloid phenotype allows immune escape of MLL-rearranged B-ALL from CD19 CAR-T-cell therapy. *Blood*, *127*.

Gaspar-Maia, A., Alajem, A., Meshorer, E., & Ramalho-Santos, M. (2011). Open chromatin in pluripotency and reprogramming. *Nat Rev Mol Cell Biol*, *12*. https://doi.org/10.1038/nrm3036

Gasper, W. C., Marinov, G. K., Pauli-Behn, F., Scott, M. T., Newberry, K., DeSalvo, G., Ou, S., Myers, R. M., Vielmetter, J., & Wold, B. J. (2014). Fully automated high-throughput chromatin immunoprecipitation for ChIP-seq: Identifying ChIP-quality p300 monoclonal antibodies. *Sci Rep*, *4*. https://doi.org/10.1038/srep05152

Goh, W. W. B., Wang, W., & Wong, L. (2017). Why batch effects matter in omics data, and how to avoid them. *Trends Biotechnol*, *35*. https://doi.org/10.1016/j.tibtech.2017.02.012

Gopalan, S., Wang, Y., Harper, N. W., Garber, M., & Fazzio, T. G. (2021). Simultaneous profiling of multiple chromatin proteins in the same cells. *Molecular Cell*, *81*(22), 4736-4746.e5. https://doi.org/10.1016/j.molcel.2021.09.019

Gorkin, D. U. (2020). An atlas of dynamic chromatin landscapes in mouse fetal development. *Nature*, *583*.

Gottgens, B. (2015). Regulatory network control of blood stem cells. *Blood*, *125*. https://doi.org/10.1182/blood-2014-08-570226

Granja, J. M. (2019). Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nat. Biotechnol.*, *37*. https://doi.org/10.1038/s41587-019-0332-7

Granja, J. M. (2021). ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.*, *53*. https://doi.org/10.1038/s41588-021-00790-6

Guenther, M. G. (2008). Aberrant chromatin at genes encoding stem cell regulators in human mixed-lineage leukemia. *Genes Dev.*, *22*.

Hainer, S. J., Boskovic, A., Rando, O. J., & Fazzio, T. G. (2018). Profiling of pluripotency factors in individual stem cells and early embryos. *BioRxiv*. https://doi.org/10.1101/286351

Harada, A. et al. A chromatin integration labelling method enables epigenomic profiling with lower input. Nat. Cell Biol.21, 287–296 (2018). (n.d.).

Hawkins, R. D., Hon, G. C., Lee, L. K., Ngo, Q., Lister, R., Pelizzola, M., Edsall, L. E., Kuan, S., Luu, Y., & Klugman, S. (2010). Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell Stem Cell*, *6*. https://doi.org/10.1016/j.stem.2010.03.018

Heaton, H. (2020). Souporcell: Robust clustering of single-cell RNA-seq data by genotype without reference genotypes. *Nat. Methods*, *17*. https://doi.org/10.1038/s41592-020-0820-1

Heintzman, N. D., Hon, G. C., Hawkins, R. D., Kheradpour, P., Stark, A., Harp, L. F., Ye, Z., Lee, L. K., Stuart, R. K., & Ching, C. W. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, *459*. https://doi.org/10.1038/nature07829

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., Barrera, L. O., Calcar, S., Qu, C., & Ching, K. A. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet*, *39*. https://doi.org/10.1038/ng1966

Heinz, S., Romanoski, C. E., Benner, C., & Glass, C. K. (2015). The selection and function of cell type-specific enhancers. *Nat Rev Mol Cell Biol*, *16*. https://doi.org/10.1038/nrm3949

Henikoff, S., Henikoff, J. G., & Ahmad, K. (2021). Simplified epigenome profiling using antibody-tethered tagmentation. *Bio. Protoc.*, *11*.

Hsieh, J. J., Cheng, E. H., & Korsmeyer, S. J. (2003). Taspase1: A threonine aspartase required for cleavage of MLL and proper HOX gene expression. *Cell*, *115*.

Hsieh, J. J., Ernst, P., Erdjument-Bromage, H., Tempst, P., & Korsmeyer, S. J. (2003). Proteolytic cleavage of MLL generates a complex of N- and C-terminal fragments that confers protein stability and subnuclear localization. *Mol. Cell. Biol.*, *23*.

Hu, D., & Shilatifard, A. (2016). Epigenetics of hematopoiesis and hematological malignancies. *Genes Dev*, *30*. https://doi.org/10.1101/gad.284109.116

Initial sequencing and analysis of the human genome | Nature. (n.d.). Retrieved January 4, 2022, from https://www.nature.com/articles/35057062

Itskovich, S. S. (2020). MBNL1 regulates essential alternative RNA splicing patterns in MLL-rearranged leukemia. *Nat. Commun.*, *11*.

Izaguirre-Carbonell, J. (2019). Critical role of Jumonji domain of JMJD1C in MLL-rearranged leukemia. *Blood Adv.*, *3*.

Jain, D., Baldi, S., Zabel, A., Straub, T., & Becker, P. B. (2015). Active promoters give rise to false positive 'Phantom Peaks' in ChIP-seq experiments. *Nucleic Acids Res*, *43*. https://doi.org/10.1093/nar/gkv637

Janssens, D. H. (2018). Automated in situ profiling of chromatin modifications resolves cell types and gene regulatory programs. *Epigenetics Chromatin*, *11*. https://doi.org/10.1186/s13072-018-0243-8

Janssens, D. H., Meers, M. P., Wu, S. J., Babaeva, E., Meshinchi, S., Sarthy, J. F., Ahmad, K., & Henikoff, S. (2021). Automated CUT&Tag profiling of chromatin heterogeneity in mixed-lineage leukemia. *Nature Genetics*, *53*(11), 1586–1596. https://doi.org/10.1038/s41588-021-00941-9

Janssens, D. H., Otto, D. J., Meers, M. P., Setty, M., Ahmad, K., & Henikoff, S. (2021). *Simultaneous CUT&Tag profiling of the accessible and silenced regulome in single cells* (p. 2021.12.19.473377). https://doi.org/10.1101/2021.12.19.473377

Karlic, R., Chung, H. R., Lasserre, J., Vlahovicek, K., & Vingron, M. (2010). Histone modification levels are predictive for gene expression. *Proc Natl Acad Sci USA*, *107*. https://doi.org/10.1073/pnas.0909344107

Kasinathan, S., Orsi, G. A., Zentner, G. E., Ahmad, K., & Henikoff, S. (2014). High-resolution mapping of transcription factor binding sites on native chromatin. *Nat. Methods*, *11*. https://doi.org/10.1038/nmeth.2766

Kaya-Okur, H. S. (2019). CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat. Commun.*, *10*. https://doi.org/10.1038/s41467-019-09982-5

Kaya-Okur, H. S., Janssens, D. H., Henikoff, J. G., Ahmad, K. & Henikoff, S. Efficient low-cost chromatin profiling with CUT&Tag. Nat. Protoc.15, 3264–3283 (2020). (n.d.).

Keene, M. A., & Elgin, S. C. R. (1981). Micrococcal nuclease as a probe of DNA sequence organization and chromatin structure. *Cell*, *27*(1, Part 2), 57–64. https://doi.org/10.1016/0092-8674(81)90360-3

Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S., & Karolchik, D. (2010). BigWig and BigBed:

Enabling browsing of large distributed datasets. *Bioinformatics*, *26*. https://doi.org/10.1093/bioinformatics/btq351
Kerry, J. (2017). MLL–AF4 spreading identifies binding sites that are distinct from super-enhancers and that govern sensitivity to DOT1L inhibition in leukemia. *Cell Rep.*, *18*.
Kidder, B. L., Hu, G., & Zhao, K. (2011). ChIP-Seq: Technical Considerations for Obtaining High Quality Data. *Nature Immunology*, *12*(10), 918–922. https://doi.org/10.1038/ni.2117
Klemm, S. L., Shipony, Z., & Greenleaf, W. J. (2019). Chromatin accessibility and the regulatory epigenome. *Nat. Rev. Genet.*, *20*. https://doi.org/10.1038/s41576-018-0089-8
Ku, W. L., Nakamura, K., Gao, W., Cui, K., Hu, G., Tang, Q., Ni, B., & Zhao, K. (2019). Single-cell chromatin immunocleavage sequencing (scChIC-seq) to profile histone modification. *Nature Methods*, *16*(4), 323–325. https://doi.org/10.1038/s41592-019-0361-7
Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., Vallejos, C. A., Campbell, K. R., Beerenwinkel, N., Mahfouz, A., Pinello, L., Skums, P., Stamatakis, A., Attolini, C. S.-O., Aparicio, S., Baaijens, J., Balvert, M., Barbanson, B. de, Cappuccio, A., … Schönhuth, A. (2020). Eleven grand challenges in single-cell data science. *Genome Biology*, *21*(1), 31. https://doi.org/10.1186/s13059-020-1926-6
Lai, W. K. M., & Pugh, B. F. (2017). Understanding nucleosome dynamics and their links to gene expression and DNA replication. *Nature Reviews Molecular Cell Biology*, *18*(9), 548–562. https://doi.org/10.1038/nrm.2017.47
Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T. R., & Weirauch, M. T. (2018). The human transcription factors. *Cell*, *172*. https://doi.org/10.1016/j.cell.2018.01.029
Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B. E., Bickel, P., Brown, J. B., & Cayting, P. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res*, *22*. https://doi.org/10.1101/gr.136184.111
Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods*, *9*. https://doi.org/10.1038/nmeth.1923
Laugesen, A., & Helin, K. (2014). Chromatin repressive complexes in stem cells, development, and cancer. *Cell Stem Cell*, *14*. https://doi.org/10.1016/j.stem.2014.05.006
Lee, T. I. (2006). Control of developmental regulators by polycomb in human embryonic stem cells. *Cell*, *125*. https://doi.org/10.1016/j.cell.2006.02.043
Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., & Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet*, *11*. https://doi.org/10.1038/nrg2825
Levine, M., Cattoglio, C., & Tjian, R. (2014). Looping back to leap forward: Transcription enters a new era. *Cell*, *157*. https://doi.org/10.1016/j.cell.2014.02.009
Liau, B. B. (2016). Adaptive chromatin remodeling drives glioblastoma stem cell plasticity and drug tolerance. *Cell Stem Cell*, *20*. https://doi.org/10.1016/j.stem.2016.11.003
Liberzon, A. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, *27*. https://doi.org/10.1093/bioinformatics/btr260
Lin, C. (2010). AFF4, a component of the ELL/P-TEFb elongation complex and a shared subunit of MLL chimeras, can link transcription elongation to leukemia. *Mol. Cell*, *37*.
Lin, S. (2016). Instructive role of MLL-fusion proteins revealed by a model of t(4;11) pro-B acute lymphoblastic leukemia. *Cancer Cell*, *30*.
Ling, X., Harkness, T. A., Schultz, M. C., Fisher-Adams, G., & Grunstein, M. (1996). Yeast histone H3 and H4 amino termini are important for nucleosome assembly in vivo and in vitro: Redundant and position-independent functions in assembly but not in gene regulation. *Genes & Development*, *10*(6), 686–699. https://doi.org/10.1101/gad.10.6.686
Liu, N., Hargreaves, V. V., Zhu, Q., Kurland, J. V., Hong, J., Kim, W., Sher, F., Macias-Trevino, C.,

Rogers, J. M., & Kurita, R. (2018). Direct promoter repression by BCL11A controls the fetal to adult hemoglobin switch. *Cell*, *173*.

Liu, Q., Jiang, C., Xu, J., Zhao, M. T., Bortle, K., Cheng, X., Wang, G., Chang, H. Y., Wu, J. C., & Snyder, M. P. (2017). Genome-wide temporal profiling of transcriptome and open chromatin of early cardiomyocyte differentiation derived from hiPSCs and hESCs. *Circ Res*, *121*. https://doi.org/10.1161/CIRCRESAHA.116.310456

Liu, X. (2017). In situ capture of chromatin interactions by biotinylated dCas9. *Cell*, *170*. https://doi.org/10.1016/j.cell.2017.08.003

Llorens-Bobadilla, E. (2015). Single-cell transcriptomics reveals a population of dormant neural stem cells that become activated upon brain injury. *Cell Stem Cell*, *17*. https://doi.org/10.1016/j.stem.2015.07.002

Luger, K., Dechassa, M. L., & Tremethick, D. J. (2012). New insights into nucleosome and chromatin structure: An ordered state or a disordered affair? *Nature Reviews Molecular Cell Biology*, *13*(7), 436–447. https://doi.org/10.1038/nrm3382

Mackay, A., Burford, A., Carvalho, D., Izquierdo, E., Fazal-Salom, J., Taylor, K. R., Bjerke, L., Clarke, M., Vinci, M., & Nandhabalan, M. (2017). Integrated molecular meta-analysis of 1000 pediatric high-grade and diffuse intrinsic pontine glioma. *Cancer Cell*, *32*.

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., & McCarroll, S. A. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, *161*(5), 1202–1214. https://doi.org/10.1016/j.cell.2015.05.002

Maethner, E. (2013). MLL–ENL inhibits Polycomb repressive complex 1 to achieve efficient transformation of hematopoietic cells. *Cell Rep.*, *3*.

Malaney, P., Nicosia, S. V., & Dave, V. (2014). One mouse, one patient paradigm: New avatars of personalized cancer therapy. *Cancer Lett*, *344*. https://doi.org/10.1016/j.canlet.2013.10.010

Martello, G., & Smith, A. (2014). The nature of embryonic stem cells. *Annu Rev Cell Dev Biol*, *30*. https://doi.org/10.1146/annurev-cellbio-100913-013116

Massague, J., & Chen, Y. G. (2000). Controlling TGF-beta signaling. *Genes Dev*, *14*.

Meers, M. P., Janssens, D. H., & Henikoff, S. (2019). Pioneer factor–nucleosome binding events during differentiation are motif encoded. *Mol. Cell*, *75*. https://doi.org/10.1016/j.molcel.2019.05.025

Meers, M. P., Janssens, D. H., & Henikoff, S. (2021). *Multifactorial chromatin regulatory landscapes at single cell resolution* (p. 2021.07.08.451691). https://doi.org/10.1101/2021.07.08.451691

Meers, M. P., Tenenbaum, D., & Henikoff, S. (2019). Peak calling by sparse enrichment analysis for CUT&RUN chromatin profiling. *Epigenetics Chromatin*, *12*.

Meissner, A. (2008). Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, *454*. https://doi.org/10.1038/nature07107

Method of the Year 2013 | Nature Methods. (n.d.). Retrieved January 4, 2022, from https://www.nature.com/articles/nmeth.2801

Mezger, A. et al. High-throughput chromatin accessibility profiling at single-cell resolution. Nat. Commun.9, 3647 (2018). (n.d.).

Milne, T. A. (2002). MLL targets SET domain methyltransferase activity to Hox gene promoters. *Mol. Cell*, *10*.

Monroe, S. C. (2011). MLL–AF9 and MLL–ENL alter the dynamic association of transcriptional regulators with genes critical for leukemia. *Exp. Hematol.*, *39*.

Morey, M., Fernández-Marmiesse, A., Castiñeiras, D., Fraga, J. M., Couce, M. L., & Cocho, J. A. (2013). A glimpse into past, present, and future DNA sequencing. *Molecular Genetics and Metabolism*, *110*(1), 3–24. https://doi.org/10.1016/j.ymgme.2013.04.024

Nagaraja, S., Vitanza, N. A., Woo, P. J., Taylor, K. R., Liu, F., Zhang, L., Li, M., Meng, W., Ponnuswami,

A., & Sun, W. (2017). Transcriptional dependencies in diffuse intrinsic pontine glioma. *Cancer Cell*, *31*.

Narendra, V., Rocha, P. P., An, D., Raviram, R., Skok, J. A., Mazzoni, E. O., & Reinberg, D. (2015). CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. *Science*, *347*. https://doi.org/10.1126/science.1262088

Neiman, M., Sundling, S., Gronberg, H., Hall, P., Czene, K., Lindberg, J., & Klevebring, D. (2012). Library preparation and multiplex capture for massive parallel sequencing applications made efficient and easy. *PLoS ONE*, *7*. https://doi.org/10.1371/journal.pone.0048616

Okada, Y. (2005). HDOT1L links histone methylation to leukemogenesis. *Cell*, *121*.

O'Neill, K. M. (2018). Depletion of DNMT1 in differentiated human cells highlights key classes of sensitive genes and an interplay with polycomb repression. *Epigenetics Chromatin*, *11*. https://doi.org/10.1186/s13072-018-0182-4

Park, D., Lee, Y., Bhupindersingh, G., & Iyer, V. R. (2013). Widespread misinterpretable ChIP-seq bias in yeast. *PLoS ONE*, *8*. https://doi.org/10.1371/journal.pone.0083506

Park, P. J. (2009). ChIP–seq: Advantages and challenges of a maturing technology. *Nature Reviews Genetics*, *10*(10), 669–680. https://doi.org/10.1038/nrg2641

Patel, A. P. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, *344*. https://doi.org/10.1126/science.1254257

Picelli, S. (2014). Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res.*, *24*. https://doi.org/10.1101/gr.177881.114

Picelli, S., Faridani, O. R., Björklund, A. K., Winberg, G., Sagasser, S., & Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. *Nature Protocols*, *9*(1), 171–181. https://doi.org/10.1038/nprot.2014.006

Policastro, R. A., & Zentner, G. E. (2018). Enzymatic methods for genome-wide profiling of protein binding sites. *Brief. Funct. Genom.*, *17*.

Pollack, J. R., Perou, C. M., Alizadeh, A. A., Eisen, M. B., Pergamenschikov, A., Williams, C. F., Jeffrey, S. S., Botstein, D., & Brown, P. O. (1999). Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nature Genetics*, *23*(1), 41–46. https://doi.org/10.1038/12640

Prange, K. H. M. (2017). MLL–AF9 and MLL–AF4 oncofusion proteins bind a distinct enhancer repertoire and target the RUNX1 program in 11q23 acute myeloid leukemia. *Oncogene*, *36*.

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*. https://doi.org/10.1093/bioinformatics/btq033

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S. A., Flynn, R. A., & Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*, *470*. https://doi.org/10.1038/nature09692

Ramirez, F., Ryan, D. P., Gruning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., Heyne, S., Dundar, F., & Manke, T. (2016). deepTools2: A next generation web server for deep-sequencing data analysis. *Nucleic Acids Res*, *44*. https://doi.org/10.1093/nar/gkw257

Rayes, A., McMasters, R. L., & O'Brien, M. M. (2016). Lineage switch in MLL-rearranged infant leukemia following CD19-directed therapy. *Pediatr. Blood Cancer*, *63*.

Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., & Clatworthy, M. (2017). The human cell atlas. *Elife*, *6*. https://doi.org/10.7554/eLife.27041

Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T. L., Wilson, C. J., Bell, S. P., & Young, R. A. (2000). Genome-Wide Location and Function of DNA Binding Proteins. *Science*, *290*(5500), 2306–2309. https://doi.org/10.1126/science.290.5500.2306

Reuter, J. A., Spacek, D., & Snyder, M. P. (2015). High-Throughput Sequencing Technologies. *Molecular Cell*, *58*(4), 586–597. https://doi.org/10.1016/j.molcel.2015.05.004

Reznikoff, W. S. (2003). Tn5 as a model for understanding DNA transposition. *Mol. Microbiol.*, *47*. https://doi.org/10.1046/j.1365-2958.2003.03382.x

Rhee, H. S., & Pugh, B. F. (2011). Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell*, *147*. https://doi.org/10.1016/j.cell.2011.11.013

Rheinbay, E. (2013). An aberrant transcription factor network essential for Wnt signaling and stem cell maintenance in glioblastoma. *Cell Rep.*, *3*. https://doi.org/10.1016/j.celrep.2013.04.021

Rosenberg, A. B., Roco, C. M., Muscat, R. A., Kuchina, A., Sample, P., Yao, Z., Graybuck, L. T., Peeler, D. J., Mukherjee, S., Chen, W., Pun, S. H., Sellers, D. L., Tasic, B., & Seelig, G. (2018). Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science (New York, N.Y.)*, *360*(6385), 176–182. https://doi.org/10.1126/science.aam8999

Rotem, A. (2015). Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol.*, *33*. https://doi.org/10.1038/nbt.3383

Rozenblatt-Rosen, O., Stubbington, M. J. T., Regev, A., & Teichmann, S. A. (2017). The human cell atlas: From vision to reality. *Nature*, *550*. https://doi.org/10.1038/550451a

Schmid, M., Durussel, T., & Laemmli, U. K. (2004a). ChIC and ChEC: genomic mapping of chromatin proteins. *Mol. Cell*, *16*.

Schmid, M., Durussel, T., & Laemmli, U. K. (2004b). ChIC and ChEC: Genomic Mapping of Chromatin Proteins. *Molecular Cell*, *16*(1), 147–157. https://doi.org/10.1016/j.molcel.2004.09.007

Schwartzentruber, J., Korshunov, A., Liu, X. Y., Jones, D. T., Pfaff, E., Jacob, K., Sturm, D., Fontebasso, A. M., Quang, D. A., & Tonjes, M. (2012). Driver mutations in histone H3.3 and chromatin remodelling genes in paediatric glioblastoma. *Nature*, *482*. https://doi.org/10.1038/nature10833

Segerman, A. (2016). Clonal variation in drug and radiation response among glioma-initiating cells is linked to proneural–mesenchymal transition. *Cell Rep.*, *17*. https://doi.org/10.1016/j.celrep.2016.11.056

Sergushichev, A. A. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. Preprint at bioRxivhttps://doi.org/10.1101/060012 (2016). (n.d.).

Skene, P. J., Henikoff, J. G., & Henikoff, S. (2018). Targeted in situ genome-wide profiling with high efficiency for low cell numbers. *Nat Protoc*, *13*. https://doi.org/10.1038/nprot.2018.015

Skene, P. J., & Henikoff, S. (2017). An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *Elife*, *6*. https://doi.org/10.7554/eLife.21856

Slany, R. K. (2016). The molecular mechanics of mixed lineage leukemia. *Oncogene*, *35*.

Sparmann, A., & Lohuizen, M. (2006). Polycomb silencers control cell fate, development and cancer. *Nat. Rev. Cancer*, *6*. https://doi.org/10.1038/nrc1991

Steensel, B. van, & Henikoff, S. (2000). Identification of in vivo DNA targets of chromatin proteins using tethered Dam methyltransferase. *Nature Biotechnology*, *18*(4), 424–428. https://doi.org/10.1038/74487

Steiniger, M., Adams, C. D., Marko, J. F., & Reznikoff, W. S. (2006). Defining characteristics of Tn5 transposase non-specific DNA binding. *Nucleic Acids Res.*, *34*. https://doi.org/10.1093/nar/gkl179

Stuart, T. (2019). Comprehensive integration of single-cell data. *Cell*, *177*. https://doi.org/10.1016/j.cell.2019.05.031

Svensson, V., Vento-Tormo, R., & Teichmann, S. A. (2018). Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.*, *13*. https://doi.org/10.1038/nprot.2017.149

Tan, J. (2011). CBX8, a Polycomb group protein, is essential for MLL–AF9-induced leukemogenesis. *Cancer Cell*, *20*.

Tanay, A., & Regev, A. (2017). Scaling single-cell genomics from phenomenology to mechanism. *Nature*, *541*. https://doi.org/10.1038/nature21350

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., Lao, K., & Surani, M. A. (2009). MRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, *6*(5), 377–382. https://doi.org/10.1038/nmeth.1315

Tedesco, M., Giannese, F., Lazarević, D., Giansanti, V., Rosano, D., Monzani, S., Catalano, I., Grassi, E., Zanella, E. R., Botrugno, O. A., Morelli, L., Panina Bordignon, P., Caravagna, G., Bertotti, A., Martino, G., Aldrighetti, L., Pasqualato, S., Trusolino, L., Cittaro, D., & Tonon, G. (2021). Chromatin Velocity reveals epigenetic dynamics by single-cell profiling of heterochromatin and euchromatin. *Nature Biotechnology*, 1–10. https://doi.org/10.1038/s41587-021-01031-1

Teytelman, L., Thurtle, D. M., Rine, J., & Oudenaarden, A. (2013). Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. *Proc Natl Acad Sci USA*, *110*. https://doi.org/10.1073/pnas.1316064110

Trapnell, C. (2015). Defining cell types and states with single-cell genomics. *Genome Research*, *25*(10), 1491–1498. https://doi.org/10.1101/gr.190595.115

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., & Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, *32*(4), 381–386. https://doi.org/10.1038/nbt.2859

Wallerman, O., Nord, H., Bysani, M., Borghini, L., & Wadelius, C. (2015). LobChIP: from cells to sequencing ready ChIP libraries in a single day. *Epigenetics Chromatin*, *8*. https://doi.org/10.1186/s13072-015-0017-5

Wang, Q. (2017). Tumor evolution of glioma-intrinsic gene expression subtypes associates with immunological changes in the microenvironment. *Cancer Cell*, *32*. https://doi.org/10.1016/j.ccell.2017.06.003

Winters, A. C., & Bernt, K. M. (2017). MLL-rearranged leukemias—An update on science and clinical approaches. *Front. Pediatr.*, *5*.

Wong, P., Iwasaki, M., Somervaille, T. C., So, C. W., & Cleary, M. L. (2007). Meis1 is an essential and rate-limiting regulator of MLL leukemia stem cell potential. *Genes Dev.*, *21*.

Wu, A. R., Neff, N. F., Kalisky, T., Dalerba, P., Treutlein, B., Rothenberg, M. E., Mburu, F. M., Mantalas, G. L., Sim, S., Clarke, M. F., & Quake, S. R. (2014). Quantitative assessment of single-cell RNA-sequencing methods. *Nature Methods*, *11*(1), 41–46. https://doi.org/10.1038/nmeth.2694

Wu, S. J., Furlan, S. N., Mihalas, A. B., Kaya-Okur, H. S., Feroze, A. H., Emerson, S. N., Zheng, Y., Carson, K., Cimino, P. J., Keene, C. D., Sarthy, J. F., Gottardo, R., Ahmad, K., Henikoff, S., & Patel, A. P. (2021). Single-cell CUT&Tag analysis of chromatin modifications in differentiation and tumor progression. *Nature Biotechnology*, *39*(7), 819–824. https://doi.org/10.1038/s41587-021-00865-z

Yoshida, H. (2019). The cis-regulatory atlas of the mouse immune system. *Cell*, *176*.

Zaborowska, J., Egloff, S., & Murphy, S. (2016). The pol II CTD: new twists in the tail. *Nat. Struct. Mol. Biol.*, *23*. https://doi.org/10.1038/nsmb.3285

Zeisig, B. B. (2004). Hoxa9 and Meis1 are key targets for MLL–ENL-mediated cellular immortalization. *Mol. Cell. Biol.*, *24*.

Zeisig, D. T. (2005). The eleven-nineteen-leukemia protein ENL connects nuclear MLL fusion partners with chromatin. *Oncogene*, *24*.

Zeleznik-Le, N. J., Harden, A. M., & Rowley, J. D. (1994). 11q23 translocations split the “AT-hook” cruciform DNA-binding region and the transcriptional repression domain from the activation domain of the mixed-lineage leukemia (MLL) gene. *Proc. Natl Acad. Sci. USA*, *91*.

Zentner, G. E., & Henikoff, S. (2014). High-resolution digital profiling of the epigenome. *Nat. Rev. Genet.*, *15*. https://doi.org/10.1038/nrg3798

Zentner, G. E., Kasinathan, S., Xin, B., Rohs, R., & Henikoff, S. (2015). ChEC-seq kinetics discriminates transcription factor binding sites by DNA sequence and shape in vivo. *Nat. Commun.*, *6*. https://doi.org/10.1038/ncomms9733

Zhang, B., Srivastava, A., Mimitou, E., Stuart, T., Raimondi, I., Hao, Y., Smibert, P., & Satija, R. (2021). *Characterizing cellular heterogeneity in chromatin state with scCUT&Tag-pro* (p.

2021.09.13.460120). https://doi.org/10.1101/2021.09.13.460120
Zhang, J. (2020). An integrative ENCODE resource for cancer genomics. *Nat. Commun.*, *11*.
Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., & Liu, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol*, *9*. https://doi.org/10.1186/gb-2008-9-9-r137
Zhao, J., Kennedy, B. K., Lawrence, B. D., Barbie, D. A., Matera, A. G., Fletcher, J. A., & Harlow, E. (2000). NPAT links cyclin E-Cdk2 to the regulation of replication-dependent histone gene transcription. *Genes Dev*, *14*. https://doi.org/10.1101/gad.827700
Zheng, G. X. (2017). Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, *8*. https://doi.org/10.1038/ncomms14049
Zhu, C., Zhang, Y., Li, Y. E., Lucero, J., Behrens, M. M., & Ren, B. (2021). Joint profiling of histone modifications and transcriptome in single cells from mouse brain. *Nature Methods*, *18*(3), 283–292. https://doi.org/10.1038/s41592-021-01060-3