Title

Stone soup: digesting the epigenome through the window of chromatin accessibility

Author

Morrow, Alyssa K

Publication Date

2021

Peer reviewed|Thesis/dissertation

Stone soup: digesting the epigenome through the window of chromatin accessibility

by

Alyssa K Morrow

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Nir Yosef, Co-chair
Professor Anthony D. Joseph, Co-chair
Professor Ellen Robey

Summer 2021

Stone soup: digesting the epigenome through the window of chromatin accessibility

Abstract

Stone soup: digesting the epigenome through the window of chromatin accessibility

by

Alyssa K Morrow

Doctor of Philosophy in Computer Science

University of California, Berkeley

Professor Nir Yosef, Co-chair

Professor Anthony D. Joseph, Co-chair

In this thesis, we leverage epigenetic datasets that provide a limited view of the epigenome to understand changes in transcriptional activity across a diverse set of cellular contexts. It is often the case that we would like to learn about an epigenome that is partially characterized by different types of experiments. In our case, the epigenome is often partially characterized using assays that measure chromatin accessibility. In the European fable of *Stone Soup*, a hungry traveler convinces a village to give them ingredients to make soup; however, all the traveler has available to them are stones. Thus, the stones provide a basis for the soup, as it brings in other crucial ingredients to design a hearty meal. In this thesis, we leverage chromatin accessibility as our stone: the primary ingredient that we leverage to gather a more complete view of the epigenome.

In the first part of this work, we explore methods designed to increase our understanding of the epigenome, primarily through the identification of genomic locations that are crucial to regulatory activity, including binding sites of transcription factors and modification of histone tails. In particular, we introduce two methods designed to predict such epigenetic events. While the first of these methods is designed to predict epigenetic events from DNA sequence alone, the second method improves upon existing methodology to predict epigenetic events in novel cellular contexts using chromatin accessibility as the primary indicator of context specificity.

In the second part of this work, we leverage chromatin accessibility, as well as additional transcriptomic epigenetic information, to characterize changes between cellular contexts for various applications. In the first application, we leverage chromatin accessibility and transcriptomics to understand how CD8(+) T cells change after reprogramming cells to a pluripotent state, and how maintenance of the T cell phenotype can ultimately effect outcomes for cancer immunotherapy. Finally, we explore the role of the transcription factor, MORC3, in

modulation of the accessible genome in the innate immune system. Together, these applications demonstrate how chromatin accessibility can be leveraged to characterize changes in regulatory activity and gain a more complete understanding of the epigenome.

To Mark Schneider, who has always been there for me, even from 5,659 miles away.

Contents

List of Figures

List of Tables

Acknowledgments

Over the course of my life, I have had the fortune of working with, living with, and being loved by an amazing group of people. I can only briefly acknowledge many of these people on this short page.

To Mark, and my immediate family (Cheryl Kramer, James Morrow, Matt Morrow, Elizabeth Morrow), who have always been there for me, and always made time to visit me in Berkeley, despite their busy lives.

To my extended family: my Grandpa Morrow and Grandma Kramer, Aunt Lori, Uncle Jeff, and everyone else for making coming home for the holidays special. For Grandpa Kramer, for being one of the only people in my life who continued to believe that getting a PhD was a good idea. For Grandma Morrow, who I wish could have been here to see me graduate.

To my amazing advisors: Nir was always there to provide detailed feedback and help me stay on track. Anthony always helped me remember the big picture, especially when I was down in the weeds.

To Dave Patterson and David Culler, who got me off the ground in Berkeley and helped me feel like I belonged there.

To Nilabh Shastri and Bill Bolosky, the first people I worked with that convinced me I was capable of doing good research.

To Jennifer Listgarten, my first real research role model.

To all my amazing Wisconsin friends who are always there to make me laugh. To Molly and Joanna (JAM) and Camp Malejajah.

To all my friends that I made in Berkeley. Regina, Hani, Esther, Vaishaal, Robert, Jeff, Ahmed, Kelly, Karl, Utkarsha, Zoe, Arya, Coline, Ben. Special shout out to Vaishaal for doing research with me as well as being a great friend.

To Devin and Frank, who were great friends and mentors throughout my PhD.

To all the amazing support from students, faculty, and staff in the AMPLab, RISELab, and Yosef Lab. To Jon and Shane for being incredible support staff. To James Kaminski, who taught me that brilliance is always best paired with modesty.

To all my colleagues I have collaborated with who work incredibly hard to run experiments in the "real" lab: Moritz, Russell Vance, Robin, Jeremy, Dharmesh, and Bruce Blazar.

To all the amazing undergraduate and masters students I have had the fortune of working with: Weston, Jahnavi, Gunjan, George, Alex, Doris, Abhishek, and Rohan.

Thanks to Hector Roux de Bézieux for help with constructing pipelines discussed in Section 3.1.

To my dogs: Riley, Maizy, and Pepper.

Chapter 1

Introduction

1.1 The epigenome explained

Many of us are quite familiar with the central dogma of biology: DNA makes RNA makes proteins. However, one critical question that is not addressed in this simplistic but catchy dogma is how this process can differ across cellular contexts. We all know that brain cells are different from skin cells, which are different from liver cells. Although these cells have similar DNA that is inherited from the germ line, the selection and strength of genes that are transcribed to RNA varies greatly from cellular context to context. One primary driver of changes in expression of genes is due to changes in the **epigenome**.

In the name "epigenome", the prefix "epi" is Greek, and can be translated to "above" [228]. Thus, the epigenome can almost literally be translated to "above the genome", and includes all modifications that control the regulation of gene expression, but do not modify DNA sequence. We define the epigenome as all events, marks, and conformational changes that affect the expression of genes.

With the advent of next generation sequencing, measuring different aspects that characterize the epigenome has never been easier. Through the utilization of different antibodies and enzymes, one can treat a sample using various protocols to obtain measurements for many epigenetic events. In this section, we will explain three types of epigenetic events, including changes in chromatin accessibility, transcription factor activity, and the modification of histone tails. These events are visualized in Figure 1.1. Although these three types of epigenetic events are not exhaustive, we focus on them because they are utilized for the development of methods and explanation of applications in this body of work.

1.1.1 Transcription factors: the main regulators of gene expression

Transcription factors (TF) are proteins that regulate the transcription of genes. For this reason, TFs have been referred to as the "main regulators of transcription" [217]. TFs

can either activate or repress the transcription of a gene, and thus regulate the amount of messenger RNA (mRNA) that is transcribed from a regulated gene [216]. TFs often bind directly to DNA sequences in promoter regions, located upstream of a gene. Here, these TFs can recruit co-activators that modify the chromatin environment and facilitate assembly of the pre-initialization complex (PIC), composed of general TFs [3] and RNA polymerase II [79]. TFs can also bind to regions distal to the affected gene. This general group of distal regions is referred to as gene regulatory regions. Within this group, a subclass of regions, called enhancers, positively regulate a gene of interest [132]. TFs can even regulate their own expression, through binding of enhancers or promoters that ultimately affect its own transcription [132].

While some TFs are general, as they are expressed and active in many of the cells in an organism, others are specific to certain types of cells, stages of development, or tissue location [216]. Examples of widely expressed transcription factors, regardless of cell type, includes the general TFs, which are required for formation of the PIC in promoter regions [41]. Examples of general TFs include TFIIA, TFIIB, and TFIID in eukaryotes [115]. Although these TFs can be found in a majority of cellular contexts, other TFs are specific to contexts. One such example of this can be shown through the modulation of TFs in the differentiation of CD8(+) T cells. While CD8(+) T cells more differentiated than the naïve phenotype show decreasing activity of TFs such as TCF7, LEF1, and KLF7, other TFs, such as TBX21 and PRDM1 show increasing levels of expression [60].

There are many factors that determine where different TFs bind across the genome. The binding affinity of many TFs depends on DNA sequence, where changes in sequence specificity reflect the biochemical properties of the TF in question. For example, families of TFs have different binding domains, which ultimately effect which sequences these TFs bind [225]. TFs bind to short DNA sequence patterns that range in length from 4 to 12 base pairs (bp) [91]. These short patterns can be represented probabilistically, where each base pair in a pattern has a probability of being observed. This probabilistic representation of DNA sequence patterns is referred to as DNA sequence motifs, or motifs for short.

Because these motifs are small in size, they can often occur randomly in the genome. For many of these motif instances, interactions between TFs and the respective DNA region it is specific to can be weak [132]. To strengthen interactions, TFs often recruit corepressors and co-activators to stabilize complexes in regulatory regions of the genome [219]. However, not all TFs have sequence specificity. In the recruitment of corepressors and co-activators, some of the recruited TFs bind in a context dependent manner, where TFs bind to the recruiter or a composite motif [105].

Apart from DNA sequence, histone placement and accessibility of chromatin can affect TF binding. As discussed in Section 1.1.2, the modification and placement of histones can affect the ability of TF binding [132]. TF binding can indirectly affect gene expression through the displacement of histones, or the recruitment of co-activators and corepressors that modulate the modification of histones [107, 229, 172]. TF binding is also affected by general accessibility of chromatin at regulatory regions. Thurman et al. have shown that 98.2% of all TF occupied sites mapped in the ENCODE consortium [34] overlap accessible

regions of the genome. However, some TFs occupy nucleosomal sites [212], which are generally inaccessible. These TFs make up heterochromatin-bound repressive complexes [35, 189].

In general, TF binding sites can be identified genome-wide through assays such as Chromatin Immunoprecipitation followed by DNA sequencing (ChIP-seq) [167] or Cleavage Under Targets and Release Using Nuclease (CUT&RUN)[193]. Although large epigenetic data consortiums today generally use ChIP-seq as the standard method for measuring TF binding sites genome wide [34, 109], assays such as CUT&RUN support in-situ identification of binding sites, while generating less background noise and requiring approximately one tenth the sequencing depth of ChIP-seq [193]. Although these assays can in principle be used to identify and characterize the activity of regulatory regions in a cellular context of interest [167], running separate ChIP-seq experiments for each of a large number of relevant TFs is time and cost-intensive and, in some instances, unfeasible due to low input size [144]. As a result, the task of predicting the location of such epigenetic events in silico in lieu of experimental evaluation received great deal of attention [220]. These methods will be discussed in greater depth in Chapter 2.

1.1.2 Histone modifications

For the sake of simplicity, DNA is often drawn in its primary structure, namely a sequence of A's, T's, G's, and C's, where each letter represents a DNA base. However, in reality, DNA takes a complex conformation which involves molecules, primarily proteins, that control which regions of DNA are actively transcribed into genes, and which regions are repressed. One of the ways that DNA maintains this complex structure within the nucleus is by wrapping around structural units called nucleosomes. Nucleosomes are sections of DNA that are wrapped around a core of proteins, or histones. Together, DNA, histone proteins, and other molecular components make up chromatin, which is the structure in which DNA is present in a cell [107]. Each of these histones has unstructured histone "tails" which protrude from the nucleosome [39]. Although these tails are unstructured, they are able to interact with DNA to help determine the formation of higher-order chromatin [86]. Each of these histone tails can be post-translationally modified, where each modification can uniquely change conformation and affect transcription, DNA repair, condensation, and DNA replication [107].

Histone tails can be modified through post-translational modifications (PTM), and include acetylation and methylation of various amino acids along the histone tail [10]. Note that these two modifications do not encapsulate all types of PTMs, but we will focus on them in particular, as they are studied in Chapter 5. These various types of PTMs can inhibit or promote transcription. One example of PTMs which activates transcription of genes is H3K4me3. While H3K4me3 activates transcription when located near promoter regions, other PTMs, such as H3K27me3, are associated with repression of transcription [123]. The effect of other PTMs are not so straightforward: various types of histone modifications can both inhibit and promote transcription, depending on their proximity to DNA regulatory regions. For example, methylation at H3K36 has a positive effect on transcription when

located near a coding region, and a negative affect when located at the promoter [107]. Similar to TFs, histone modifications can be experimentally measured using assays such as ChIP-seq [167] and CUT&RUN [193].

Regardless of the type of PTM, histone modifications affect transcription through two main mechanisms. This includes (1) unraveling chromatin through disruption of contact with nucleosomes, or (2) recruiting non histone proteins to modify chromatin state [107]. One example in which a PTM recruits non histone proteins to change chromatin state is through the recruitment of a TF, MORC3, via H3K4me3 [67]. While the MORC family of TFs is known to play roles in gene silencing, H3K4me3 is a marker of active regions [118]. Here, it has been shown that MORC3 can aid in the silencing of genes in mouse embryonic stem (ES) cells through recruitment of MORC3 by H3K4me3 to active promoter regions [118]. We discuss this example in particular, as Chapter 5 discusses the role of MORC3 in transcriptional silencing, albeit in human cells.

1.1.3 Chromatin accessibility: the stone of the epigenome

To date, one of the most prevalent types of genomic-scale epigenetic information is chromatin accessibility. Chromatin accessibility refers to the physical compaction of chromatin, where nucleosomal compacted regions are less accessible, while nucleosomal depleted regions are more accessible. Because nucleosomes are not evenly spaced throughout the genome, regions of accessibility can be varied, and are found in different locations depending on the cellular context considered [139].

Respective assays such as DNase I hypersensitive sites sequencing (DNase-seq) [20] and Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) [24] identify accessible chromatin regions genome-wide through enzymatic cleavage of exposed DNA. DNase-seq and ATAC-seq protocols use cleavage enzymes (DNase-I and Tn5, respectively) to cleave DNA of open chromatin [119]. These cleaved segments of DNA are fragmented, then sequenced and aligned to quantify levels of accessibility genome wide [232]. Although the sensitivity and specificity of ATAC-seq is similar to DNase-seq, it requires only a fraction of the cells of DNase-seq, making it a more feasible measure of chromatin accessibility for samples with low yields [24].

One of the reasons that chromatin accessibility is well characterized is because of its wide availability. ENCODE alone contains 300 biosamples with measurements of DNase-seq. Additionally, in the advent of ATAC-seq, ENCODE has collected ATAC-seq from 48 primary tissues for human alone [141]. Unlike measurements for TFs and histone modifications, measuring chromatin accessibility requires a single experiment. However, individual TFs and PTMs must be collected as separate experiments for each target of interest. Although we could, in principle, fully characterize the activity of regulatory regions in a cellular context of interest [167], running separate experiments for each of a large number of relevant DNA binding proteins and histone modifications is unfeasible.

Besides wide availability, chromatin accessibility is valuable because of its ability to indicate regulatory regions and enhance our understanding of transcriptional regulation,

outside of accessibility alone. Generally, regions of accessibility can indicate regulatory elements [139]. However, beyond accessibility, these regulatory regions are often associated with other epigenetic events such as binding of transcription factors (TFs) and other DNA binding proteins, or chemical modifications of histone tails [212]. Often, chromatin accessibility is used to observe variation in accessibility at the protein-DNA physical interface, creating unique patterns, called footprints. These footprints not only indicate regions of accessibility, but can also be used to identify footprints for TFs and chromatin modifiers that bind in accessible regions [220, 136]. Because of this, many methods exist that leverage chromatin accessibility to predict TF binding [144, 166, 96, 186, 170, 100]. Together, this means that chromatin accessibility can be used to identify regulatory regions of the genome, as well as estimate regions of TF binding and histone modifications in a cellular context of interest.

In this thesis, we refer to chromatin accessibility as the "stone" of all epigenetic measurements, as it provides the basis for which we build on to understand changes in the epigenome across cellular contexts. We refer to chromatin accessibility as the stone of our soup, not because it provides us with the most information, but because it is readily available, especially in our applications. These applications of how we leverage chromatin accessibility in particular to understand changes in the innate and adaptive immune system are discussed in Chapters 4 and 5. Besides availability, we refer to chromatin accessibility as the "stone" because of its ability to indicate regulatory regions and enhance our understanding of transcriptional regulation, outside of accessibility alone.

Despite the wide usage of chromatin accessibility for identification of regulatory regions, chromatin accessibility alone cannot identify all regulatory regions in isolation. One such example of this is in the identification of active enhancers, for which measurements of PTMs can be more informative for their identification [83]. Because of this, we note that chromatin accessibility alone is not perfectly indicative of all regulatory regions.

1.2 Dissertation Overview

In this work, I first present methods and tools that we have conceptualized and developed for processing and learning from epigenetic datasets in new cellular contexts. In the cases we consider, the epigenome in a cellular context of interest is partially characterized by gathering experimental information that measures chromatin accessibility, histone modifications, TF binding patterns, or a combination of these. Therefore, these methods and tools are designed to extrapolate knowledge from partially characterized cellular contexts to make the most of the data that we have. In Chapter 2, we discuss two methods we have developed to predict TF binding sites genome wide. While the first method leverages DNA sequence to learn motif preferences of various TFs [143], the second method leverages chromatin accessibility, as well as histone modifications, when available, to predict binding sites of TFs and chromatin modifiers in a cellular context aware manner [144].

In Chapter 3, we discuss three tools, or extensions of existing tools, designed to simplify

processing and understanding of epigenetic datasets. The first two methods build on previous work [97] to support processing of TF binding sites and histone modifications from CUT&RUN [193] in an integrated pipeline for processing epigenetic datasets. We particularly use these pipelines in Chapter 5 to process CUT&RUN for a TF called MORC3, as well as various histone modifications, in monocytes. Our second extension allows for the quantification of endogenous retroviruses from RNA sequencing datasets. This implementation is discussed in Section 3.1, and its application is discussed in Chapter 5. Finally, we introduce a set of tools that supports visualization of genomic sequencing datasets in a python environment. These tools allow us to visualize genomic datasets in the same environment used for method development and data analysis, supporting for quick and ad-hoc visualization of data that is not necessarily stored in traditional bioinformatics file formats [145].

Finally, we discuss two applications that leverage chromatin accessibility, among other epigenetic and transcriptomic datasets, to characterize cellular contexts of interest. The first application discussed in Chapter 4 leverages chromatin accessibility and transcriptomic information to characterize phenotypic differences between naïve and stem memory CD8(+) T cells, and how reprogramming these cells changes the epigenome. The second application, discussed in Chapter 5, leverages chromatin accessibility to identify regulatory regions crucial to the function of a TF called MORC3 in human monocytes.

In Chapter 4, we aim to measure the extent to which naïve and stem memory CD8(+) T cells retain epigenetic and transcriptomic memory after reprogramming to induced pluripotent stem cells (iPSCs). We consider this application in particular due to its role in protocol for the generation of modified T cells for adoptive T cell therapy [94, 153]. To assess the extent to which CD8(+) T cells retain memory after reprogramming, we leverage measurements of chromatin accessibility to evaluate regulatory regions that maintain and lose accessibility after reprogramming, and whether retained regions are important to their respective parent phenotype. We additionally collect measurements of the transcriptome to determine which genes maintain and lose expression after reprogramming, and whether changes in accessibility nearby genes correlates to changes in gene expression for key CD8(+) associated and pluripotent associated genes.

In Chapter 5, we evaluate the role of MORC3, a TF, in the negative regulation of interferon response in the innate immune system. To evaluate the role of MORC3, we gather measurements of chromatin accessibility to isolate potential regions in which MORC3 interacts with to repress interferon response. We additionally collect measurements of the transcriptome to identify which genes MORC3 regulates. Together, these applications exemplify cases where chromatin accessibility can be leveraged to gain a more complete understanding of a cellular context of interest.

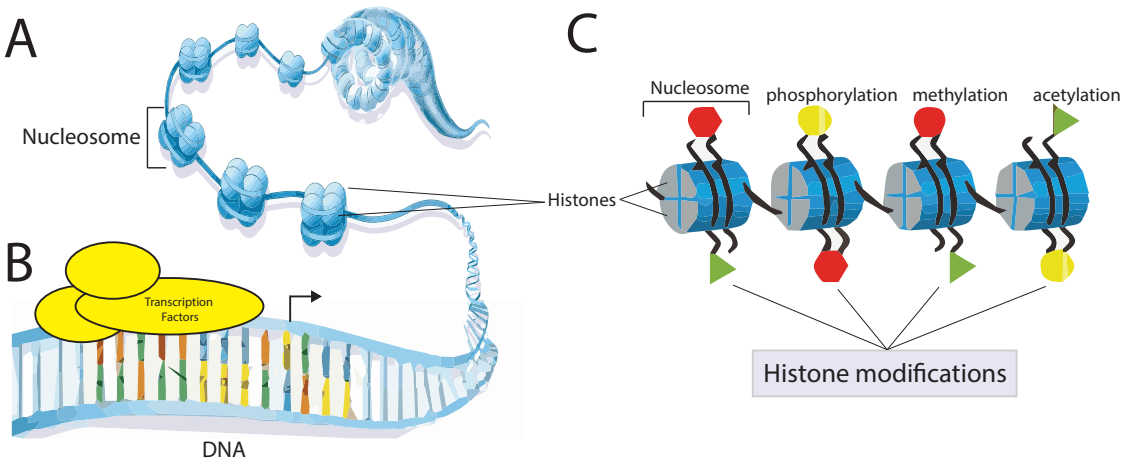


Figure 1.1: Overview of epigenetic events that affect gene expression. a) Chromatin accessibility refers to the physical compaction of chromatin, where nucleosomal compacted regions are less accessible, while nucleosomal depleted regions are more accessible. Figure is modified from Bailey et al. [9]. b) Transcription factors bind to DNA or interact with co-factors in a context dependent manner to regulate the transcription of genes. c) Nucleosomes are sections of DNA wrapped around histone proteins. Histones have unstructured histone tails which can be modified through post-translational modifications. These modifications include acetylation, methylation, and phosphorylation, for example. Figure is modified from Monteys et al. [4].

Part I

Methodology for processing and learning from epigenetic datasets

Chapter 2

Models for prediction of epigenetic events

2.1 Introduction

As discussed in Section 1.1.1, TFs are known as the main regulators of gene expression, and are often indicative of the function of their respective region [34]. Although TF binding sites (TFBS) could be evaluated genome-wide through assays such as Chromatin Immunoprecipitation followed by DNA sequencing (ChIP-seq) [167] or Cleavage Under Targets and Release Using Nuclease (CUT&RUN) [193], running separate ChIP-seq experiments for each of a large number of relevant DNA binding proteins and histone modifications is time and cost-intensive and, in some instances, unfeasible due to low input size.

As a result, the task of predicting the location of such epigenetic events in silico in lieu of experimental evaluation has received great deal of attention [220]. Methods developed to predict such events can be broadly categorized based on the genomic properties they use as features for drawing predictions. The first broadly utilized category of classification methods is restricted to using only DNA sequences as features [5, 240, 235, 190]. Although these methods cannot predict epigenetic events that are specific to cellular contexts not seen during training, they can be used to explain the effect of changes in DNA sequence on the strength of the signal of DNA binding proteins and histone modifications. In this body of work, we refer to this group of methods that use DNA sequence to predict epigenetic events as **non-generalizable**, due to the fact that they are not able to predict epigenetic events specific to a cellular context that was not observed during model training. In Section 2.2.1, we introduce a novel, non-generalizable method that uses convolutional kitchen sinks [169] to predict TFBS from DNA sequence, using fewer computational resources than traditional approaches [143].

Although non-generalizable methods inform us of the importance of DNA sequence on the locations of epigenetic events, these methods cannot generalize to new cellular contexts. This is due to the fact that non-generalizable methods do not use information collected from

cellular contexts, and thus have no way of modifying predictions based on the context under consideration. In order to consider cellular contexts when predicting epigenetic events, a second group of classification methods utilize available measurements of chromatin state from a cellular context of interest to predict epigenetic events which were not measured experimentally [166, 100, 96, 170]. We refer to these methods as **generalizable** because they are able to learn from epigenetic features from a set of contexts and be applied to predict epigenetic events in new contexts not seen during training. Many of these methods use chromatin accessibility in particular as an indication of context specificity due to its prevalence and ability to capture nuanced variation in accessibility at the protein-DNA physical interface [220, 136]. In Section 2.3.1, we introduce a novel method, Epitome, that incorporates chromatin state to predict the probability of occurrence of various epigenetic events genome wide, including TFBS and histone modifications [144]. We further demonstrate how Epitome can be extended to consider all epigenetic information available from a cellular context of interest to improve predictive accuracy of epigenetic events. Finally, in Section 2.3.2, we extend Epitome to predict epigenetic events at single cell resolution by leveraging single-cell sequencing assay for transposase-accessible chromatin (scATAC-seq) as the primary indicator of cellular context specificity.

2.2 Non-generalizable models for prediction of epigenetic events

There have been various discriminative and generative methods designed to classify short (<1kbp) DNA sequences. Here, we specifically refer to methods that classify DNA sequences based on the epigenetic events they may be associated with: namely, sequences that bind TFs. In this section, we outline numerous state-of-the-art discriminative approaches for classifying DNA sequences based on the epigenetic events they are associated with. This section is based on Morrow et al. [143], which focuses on a novel non-generalizable model for classification of DNA sequences based on the epigenetic events they are associated with.

The first group of methods we explore that have been designed to classify DNA sequence are string kernel methods. String kernel methods are well understood and have been extensively used for sequence classification [87, 50, 116]. Specifically, Fletez-Brant et al. and Lee et al. [54, 114] have applied string kernel methods to the problem of predicting TFBS. Fletez et al. [54] use the Spectrum Kernel formulated by Leslie et. al. [116] to learn the location of TFBS from DNA motifs of length n (Equation 1). We let alphabet $\mathcal{A} = \{A, T, C, G\}$ represent the set of DNA bases.

Definition 1 (Spectrum Kernel)**.** *Let $\mathcal{S}_n(\mathcal{A})$ be the set of all length n contiguous substrings in $\mathcal{A}$, and $\#(x, s)$ count the occurrences of $s \in x$ [116].*

$$K_{spec}(x, y) = \sum_{s \in \mathcal{S}_n(\mathcal{A})} \#(x, s)\#(y, s) \tag{2.1}$$

The Spectrum Kernel captures similarities between sequences when motif length l is small. However, as l increases, it becomes intractable to compare sequences due to the increased sparsity of motif presence. Specifically, for TFBS, this poses a problem when motifs range in length from 6 to 30 base pairs. The Gapped Kmer Kernel by Lee et al. [114] resolves the issue of sparsity with large l by allowing for gaps when comparing two motifs (Definition 2).

Definition 2 (GKM-Kernel)**.** *Let $\mathcal{S}_{l,k}(\mathcal{A})$ be the set of all length l contiguous substrings with k informative characters in $\mathcal{A}$, and $\#(x, s_k)$ count the occurrences of $s_k \in x$ with exactly k mismatches [114].*

$$K_{GKM}(x, y) = \sum_{s_k \in \mathcal{S}_n(\mathcal{A})} \#(x, s_k)\#(y, s_k) \tag{2.2}$$

However, kernel methods require pairwise comparison between all n training sequences and thus incur an expensive $\mathcal{O}(n^2)$ computational and memory complexity, making them computationally intractable for large data sets. To bypass $\mathcal{O}(n^2)$ memory complexity, LS-GKM SVM [113] implements the GKM Kernel using LIBSVM [27], which uses decomposition methods to iteratively solve for the SVM problem, bypassing loading the $\mathcal{O}(n^2)$ kernel matrix into memory.

As it has been repeatedly observed that the presence of a TF binding site may depend on a wider context around the site [65, 29], many approaches use convolutional neural networks (CNN) to consider wide genomic contexts around a region of interest to predict epigenetic signal [240, 5]. CNNs generalize well by encoding spatial invariance during training. Fast convolutions on a Graphical Processing Unit (GPU) allow CNNs to train on large datasets. One of the first CNNs designed to predict TFBS was DeepBind [5]. DeepBind is a convolutional neural network that trains and parameterizes a separate predictive model for each TF, based on the locations of DNA motifs for the TF of interest [5]. Model hyper parameters vary for each transcription factor, ranging from 0 to 1 hidden layers, and varying in motif length from 16 to 32.

The second, and more widely utilized neural network which predicts TF binding sites is DeepSEA. DeepSEA trains one model to predict binding for hundreds of diverse regulatory elements from DNA sequence. These regulatory elements include TFBS, regions of accessible chromatin, and the presence of histone modifications [240]. The primary difference between DeepSEA and DeepBind is that, unlike DeepBind, DeepSEA uses signal identified from ChIP-seq to indicate response variables for epigenetic events, instead of DNA motifs. This allows DeepSEA to predict epigenetic events specific to cellular contexts that it observes during training. However, because only DNA sequence is used as features, it cannot predict events in cellular contexts not seen during training. Hence, we include this method in the group of non-generalizable methods.

2.2.1 Convolutional Kitchen Sinks for Prediction of transcription factor binding sites

In this section, we present a simple and efficient non-generalizable method leveraging convolutional kitchen sinks (CKS) for predicting TFBS from DNA sequence. This text is modified from that described in Morrow et al. [143]. Unlike kernel methods, CKS bypasses the need for $\mathcal{O}(n^2)$ memory and time complexity. However, unlike CNN's, CKS train in a fraction of the time with a fraction of the parameters. Our method computes a random approximation of a convolutional kernel feature map from DNA sequence and then learns a linear model from the approximated feature map. Our method outperforms state-of-the-art DeepBind on 9 out of 10 datasets from the ENCODE consortium, while training in less than one eighth the time.

In this work, we present a convolutional kernel approximation algorithm that maintains the spatial invariance and computational efficiency of CNNs. Dubbed Convolutional Kitchen Sinks (CKS), our algorithm learns a model from the output of a 1 layer random convolutional neural network [169]. All the parameters of the network are independent and identically distributed (IID) random samples from a gaussian distribution with a specified variance. We then train a linear model on the output of this network.

The task of TFBS prediction from DNA sequence reduces to binary sequence classification. We present a randomized algorithm for finding an embedding of sequence data apt for linear classification (Algorithm 1). Our algorithm is closely related to the work of convolutional kernel networks, which approximates a convolutional kernel feature map via a nonconvex optimization objective [130]. However, unlike Mairal et al. [130], we approximate the convolutional kernel feature map via random projections in the style of Rahimi et al. [169, 168].

We will first define the convolutional n-gram kernel, and then analyze why it has desired properties for the task of string classification. Note that we use the term n-gram to refer to a contiguous sequence of n characters, whereas computational biology literature refers to the same concept as a k-mer.

Definition 3 (Convolutional n-gram kernel)**.** *Let x, y be strings of length d from an underlying alphabet $\mathcal{A}$, and let $\mathbb{H}(x, y)$ denote the Hamming distance between the two strings. Let $x_{i:j}$ denote the substring of x from index i to $j-1$. Let n be an integer less than d and let γ be a real valued positive number denoting the width of the kernel. The kernel function $K_{n,\gamma}(x, y)$ is defined as:*

$$K_{n,\gamma}(x, y) = \sum_{i=0}^{d-n} \sum_{j=0}^{d-n} \exp(-\gamma \mathbb{H}^2(x_{i:i+n}, y_{j:j+n})) \tag{2.3}$$

To gain intuition for the behavior of this kernel, take γ to be a large value. It follows that $\exp(-\gamma \mathbb{H}^2(x_{i:i+n}, y_{j:j+n})) \approx \mathbb{1}[x_{i:i+n} = y_{j:j+n}]$.

This combinatorial reformulation results in the following well studied Spectrum Kernel (Definition 1).

Algorithm 1 Convolutional Kitchen Sink for sequences

Input $x_i \dots x_N \in \mathbb{R}^d$ (input sequences), γ (width of kernel), n (convolution size), M (the approximation dimension, number of kitchen sinks)

Output $\phi(x_i) \dots \phi(x_N)$

1: **for** $j \in \{0 \dots M\}$ **do**
2: $w_j \sim \mathcal{N}(0, \gamma I_n)$ ▷ Sample kitchen sink from gaussian
3: $b_j \sim U(0, 2\pi)$ ▷ Sample phase from uniform disk
4:
5: **for** $i \in \{0 \dots N\}$ **do**
6: $z_{ij} = w_j * x_i$ ▷ Convolve filter with input sequence
7:
8: $c_{ij} = \cos(z_{ij} + b_j)$ ▷ Add phase and compute element-wise cosine
9: ▷ Note z_{ij} and c_{ij} are vectors in $\mathcal{R}^{d-n+1}$
10:
11: $\phi(x_i)_j = \sqrt{\frac{2}{M}} \sum_{k=0}^{d-n} c_{ijk}$ ▷ Average to get jth output feature value for sequence x_i
12:
13: **end for**
14: **end for**

Other string kernel methods such as the mismatch [50] and gapped n-gram kernel [63] allow for partial mismatches between n-grams. We note that decreasing γ in Equation 2.3 relaxes the penalty of n-gram mismatches between disappoints, thereby capturing the behavior of the mismatch and gapped n-gram kernels [50, 63]. Note that Equation 2.3 is computationally prohibitive, as it takes $\Omega(nd^2)$ to compute each of the N^2 entries in the kernel matrix. Furthermore, the feature map induced by the kernel in Equation 2.3 is infinite dimensional, so the kernel matrix is necessary.

Instead, we turn to a random approximation of Equation 2.3 (see Algorithm 1). Since our kernel is a sum of nonlinear functions it suffices to define a feature map $\hat{\phi}$ on sequences x and y that approximates each term in the sum from Equation 2.3:

$$\exp(-\gamma \mathbb{H}(x_{i:i+n}, y_{j:j+n})) \approx \hat{\phi}(x_{i:i+n})^T \hat{\phi}(y_{j:j+n}) \tag{2.4}$$

Claim 1 from Rahimi et al. [168] states that for $j \in \{0 \dots M-1\}$, if we choose $\hat{\phi}(x_{i:i+n})_j = \sqrt{\frac{2}{M}} \cos(w_j^T x_{i:i+n} + b_j)$, where $w_j \sim \mathcal{N}(0, \gamma)$, $b_j \sim U(0, 2\pi)$, then $\hat{\phi}(x_{i:i+n})$ satisfies Equation 2.4. Note that to use Claim 1, we represent Hamming distance in Equation 2.3 as an L2 distance. We refer to each w_j as a "random kitchen sink". The result in in Rahimi et al. [168] (Claim 1) gives strong guarantees that $\hat{\phi}(x)^T \hat{\phi}(y)$ concentrates exponentially fast to Equation 2.3, which means we can set M, the number of kitchen sinks, to be small.

Algorithm 1 details the kernel approximation. Note that in Algorithm 1, line 6 we reuse w_j across all $x_{i:i+n}$ in Equation 2.3 by a convolution. Algorithm 1 is a finite dimensional ap-

proximation of the feature map induced by the kernel in Equation 2.3 directly, circumventing the need for a kernel matrix. The computational complexity of Algorithm 1 is $\mathcal{O}(NMdn)$.

For the task of TF binding site prediction we set $n = 8$, similar to common parameter configuration for DNA sequence [54, 5, 63].

2.2.1.1 CKS's outperform standard CNNs designed to predict transcription factor binding sites

We compare our CKS to DeepBind, a state-of-the-art CNN approach for predicting transcription factor (TF) binding sites. We compare to DeepBind over other CNN methods [240, 101] due to its primary attention to DNA sequence specificity and ability to identify fine grained (101 bp) locations of binding affinity.

We train and evaluate on datasets preprocessed from the ENCODE consortium. Because binding affinity is TF specific, we use separate train and evaluation sets for each TF.

We use the same training sets as DeepBind's publicly available models. We then evaluate on DeepBind's test sets as well as a larger dataset processed directly from ENCODE.

DeepBind's test sets consist of 1000 regions for each cell type over six TFs. Each set consists of 500 positive sequences extracted from regions of high ChIP-seq signal and 500 synthetic negative sequences generated from dinucleotide shuffle of positive sequences [5].

The second test dataset consists of 100,000 regions extracted from ChIP-seq datasets for TFs ATF2 and EGR1 across multiple cell types. Positive sequences are extracted from regions of high ChIP-seq signal. Negative sequences are extracted from regions of low ChIP-seq signal with exposed chromatin.

We compare DeepBind against CKS using area under the curve (AUC) of Receiver Operating Characteristic (ROC). We choose AUC as a metric for binary classification due to its ability to measure both TF binding site detection and false positive rates.

We detail our experimental results and compare to DeepBind's pretrained models in Table 2.1. We also show ROCs for ATF2 and EGR1 on both datasets in Figure 2.1.

Our AUC is competitive (within 0.01) or superior to that of DeepBind except for ATF2 on MCF7 cell type. Furthermore on five out of six large ENCODE test sets, our AUC is strictly greater than DeepBind.

We measure DeepBind's training time on TF EGR1, trained on K562 with $72,996$ train sequences. DeepBind's training procedure takes 6497 seconds to learn 2123 parameters. For comparison, training time for CKS takes 712 seconds (Table 2.1) to learn 16,384 parameters, which is approximately eight times faster than DeepBind's runtime.

2.3 Generalizable models for prediction of epigenetic events

So far, all methods that have been discussed are not able to predict epigenetic events in a held out cellular context. This means that if we have available chromatin accessibility,

or some other measurement, collected from a cellular context of interest, there is no way of incorporating this information into the model to augment predictions based on the context in question. In this section, we discuss a class of methods that can generalize to new cell contexts by incorporating either chromatin accessibility, transcriptomics, histone modifications, or a combination thereof to inform the model of context. We refer to these types of methods as **generalizable** because they are able to train on a set of cellular contexts, then apply learned rules to predict epigenetic events in cellular contexts not seen during training. In this section, we introduce three groups of generalizable methods for predicting epigenetic signal. We then introduce a novel hybrid method, Epitome, that outperforms state-of-the-art generalizable methods for predicting TFBS and histone modifications.

The first group of generalizable methods is referred to as 'footprinters'. Footprinter methods are designed to predict epigenetic signal by utilizing available measurements of chromatin accessibility from a cellular context of interest to predict epigenetic events which were not measured experimentally [166, 96, 170, 19, 70, 31]. These methods use chromatin accessibility in particular as an indication of context specificity due to its prevalence and ability to capture nuanced variation in accessibility at the protein-DNA physical interface [220, 136]. This variation in accessibility at the protein-DNA interface forms patterns unique to TFs, and are referred to as footprints. Footprints can be distinct for different TFs in consideration, and can thus be used as features to predict TFBS [75]. These methods are considered to be generalizable because they can train using chromatin accessibility from one cellular context, and then can be applied to gather predictions in a held out context. Two subgroups of footprinters exist, and are referred to as either *de novo* [19, 70, 31] or *motif-centric* [166, 170]. De novo footprinters learn general footprints common to all TFs from a chromatin accessibility sample, then use motif databases to scan regions overlapping discovered foot-

Table 2.1: Comparison of ROC Area under Curve values (AUC) between DeepBind and CKS tested on 500 bound regions from ENCODE and 500 synthetic unbound regions.

TF	**Train Cell Type**	**Test Cell Type**	**Train Size (MB)**	**Train Time**	**DeepBind AUC**	**CKS AUC**
ATF2	H1-hESC	GM12878	10998	154s	0.72	0.77
ATF3	H1-hESC	HepG2	8616	139s	0.94	0.95
ATF3	H1-hESC	K562	8616	139s	0.83	0.84
CEBPB	HeLa-S3	A549	121010	1620s	0.99	0.99
CEBPB	HeLa-S3	K562	121010	1620s	0.99	0.98
EGR1	K562	GM12878	72996	772s	0.94	0.96
EGR1	K562	H1-hESC	72996	772s	0.87	0.92
EP300	HepG2	SK-N-SH	54828	519s	0.67	0.70
EP300	HepG2	K562	54828	519s	0.66	0.81
STAT5A	GM12878	K562	13846	199s	0.65	0.79

prints to identify which TFs bind. Motif-centric methods take a different approach by first identifying a set of candidate motif regions for each TF in question, and then learn which candidate sites overlap footprints [166]. One such example of this subgroup is DefCoM, which is a footprinter that learns patterns of Tn5 cut sites overlapping known motifs. Regardless of the subgroup, these footprinters are constrained to only identify potential TFBS in regions overlapping motifs. However, it has been shown that TFs can bind in regions not overlapping canonical motifs associated with the TF of interest [44]. We additionally demonstrate this in Figure 2.2, which shows that the ratio of ChIP-seq peaks that do not overlap any motif for a TF in question can range from 0.45 to 1.0 for 77 TFs and chromatin modifiers evaluated from ENCODE [35].

Because limiting the prediction regime to sites that overlap motifs can eliminate many TFBS, other methods have been designed to predict epigenetic events genome wide to eliminate this constraint. These methods augment the feature space to additionally incorporate DNA sequence information, RNA-sequencing information, or other epigenetic information specific to the cellular context of interest [100, 77]. One example of such methods is Catchitt [100], a co-winner of the ENCODE-DREAM TFBS prediction challenge [108], which uses a combination of features including DNA sequence content, motif hits, expression levels of transcription factors (from RNA-seq), and chromatin accessibility (DNase-seq) to predict TFBS. Similar to DeFCoM, Catchitt considers accessibility data in a quantitative form, albeit at a default binned resolution of 50bp. It also uses a wider genomic context, taking up to 2,000 bp around the candidate site of interest. Catchitt then models the joint distribution of these various different features by a simple product of independent densities or discrete distributions. Another method, called Anchor, a co-winner of the ENCODE-DREAM TFBS prediction challenge, also uses a combination of DNA sequence, chromatin accessibility, and motif information as features.

One common property of the methods designed to predict epigenetic signal in a new cellular context discussed so far is that they learn a single model that is applied similarly to all positions in the genome. In this position-agnostic approach, models are applied separately for each candidate locus, using its local properties as features (such as the occurrence of DNA binding motifs or the enzymatic cleavage patterns of chromatin accessibility). To ensure accuracy and generalizability, these models are trained to identify local properties that are commonly predictive in many different loci. The natural caveat in this approach is that different loci may largely differ in terms of which specific features (or combination thereof) are in fact predictive. For instance, a single TF can bind the genome while interacting with different factors (i.e. by co- binding or tethered binding [152]), thus leading to different footprints and, possibly, different DNA binding motifs [44, 152].

In order to solve this caveat of position-agnostic learning, a third group of methods, referred to as imputation methods, directly uses epigenetic signal from known cellular contexts to predict in a new cellular context, instead of constructing complex features [186, 48, 45]. Novel imputation methods, such as Avocado [186] and PREDICTD [45], specifically use tensor factorization, assuming a low rank representation of the feature space, and jointly learn a model for all missing epigenetic signals [186, 45]. However, a caveat of joint learning

of epigenetic signal in a single model is that objectives for each epigenetic signal can contradict each other, and update model parameters at different rates [200]. These limitations of joint learning can produce sub optimal predictions for epigenetic signal, compared to models optimized for a particular epigenetic signal of interest.

2.3.1 Epitome: Predicting epigenetic events in novel cell types with multi-cell deep ensemble learning

Considering the caveats of existing methods for predicting epigenetic signal, as discussed in Section 2.3, we present **Epitome** [144]. This section is modified from Morrow et al. [144]. Epitome is a conceptually simple alternative for predicting epigenetic events, such as TF binding and histone modifications. Similar to imputation methods, Epitome uses known epigenetic signal from multiple known cell types to predict epigenetic signal in a held out cellular context, bypassing the need to learn complex rules based on chromatin footprinting and DNA sequence. However, Epitome differs from imputation methods in three ways. First, Epitome approaches the problem of predicting epigenetic events as a classification task, treating each epigenetic signal as a binary event. As we explain in Section 2.3.1.1, this approach helps mitigate noise that arises from variation across different experiments, antibodies, and protocols that may affect quantitative results. Secondly, imputation methods such as Avocado [186] and PREDICTD [45] use a tensor factorization scheme, which assumes that epigenetic signals can be largely explained through a low dimensional representation, and looks for a single decomposition scheme for the entire tensor that couples all prediction tasks. Epitome has the flexibility of treating each prediction task completely independently from each other. As we show in Figure 2.3, Epitome performs better when considering each epigenetic signal independently. Lastly, Epitome explicitly computes local similarities between the reference cell types and the query cellular context, and uses these as features in the model. This attribute helps the model explicitly learn the importance of experimental data from reference cell types when predicting in the query cell type.

As input, Epitome requires chromatin accessibility in a query cellular context and a set of reference cell types in which the epigenetic event of interest was assayed. Epitome particularly requires chromatin accessibility as its primary indicator of cell type specificity because it is normally easy to generate [24, 20] and is informative of epigenetic events [220]. It then "copies over" epigenetic events from the reference cell types to the query cellular context in positions where their chromatin accessibility is similar. In this case, we define similarity by comparing chromatin accessibility profiles in regions of different resolution surrounding the genomic locus in question. This similarity, along with the manner by which evidence across multiple reference cell types is aggregated, are learned using a neural network.

At the heart of our approach is the reliance on large amounts of publicly available data. While it is the case that practically any given cellular context will have a uniquely characteristic epigenome, there is a great deal of overlap between contexts. Consequently, epigenetic events that are uniquely observed in one cellular context become less prevalent as the number

of other cellular contexts that have measured epigenetic events become available. This phenomenon is demonstrated in Figure 2.5, showing that for a given epigenetic event (binding of a certain TF, or a certain histone modification) the prevalence of genomic sites that are uniquely observed in only one cellular context in the ChIP-Atlas database [158] decreases substantially with the number of cellular contexts in which this event has been assayed with ChIP-seq. Particularly, we observe a mean level of coverage, or sensitivity, of over 90% for widely-assayed TFs and histone modifications (measured in 26 or more cellular contexts available in ChIP-Atlas). Looking ahead, we expect a continued increase in the size and quality of data sets in the public domain, leading to further increase in coverage.

An important caveat of Epitome is that it will miss all sites that were not observed in any other cellular context. In the following, we show that in practical application this is a reasonable compromise. We compare Epitome to various methods that are designed to predict histone modifications and protein binding sites in novel cellular contexts. These methods include TF footprinting methods [166] and methods that use a combination of features constructed from chromatin accessibility, epigenetic signal, and DNA sequence [100, 108, 186]. Regardless of the choice of features used, Epitome achieves state-of-the-art accuracy when predicting TF binding sites in held out cellular contexts and chromosomes. We additionally demonstrate the deleterious effect of joint learning for multiple epigenetic signals and suggest that Epitome, along with current imputation methods, may achieve optimal performance by using a loss function dedicated to one epigenetic signal of interest. We additionally show how Epitome can extend its definition of cell type similarity to incorporate commonly assayed histone modifications, in additional to chromatin accessibility. We show that this extension of similarity between reference cell types and the query cellular context can further improve predictive performance.

2.3.1.1 Overview of Epitome

Epitome predicts the genomic locations of epigenetic events that can be measured through ChIP-seq, including protein-DNA binding sites, histone modifications, chromatin modifier binding sites, and locations of histone variants. Epitome can predict unmeasured epigenetic events in any cellular context of interest, as long as genome-wide measurements of chromatin accessibility in that context is available. This is shown in Figure 2.6, where chromatin accessibility from a query cell type c' is compared to chromatin accessibility from reference cell types in order to predict epigenetic signal in c'. As input, Epitome requires peak called DNase-seq or ATAC-seq in the form of bed or narrow peak file formats from the query cellular context of interest, as well as a similar formatted file containing all genomic regions to be queried. For training, Epitome additionally requires a set of cell types that have both measured epigenetic signal for the epigenetic event of interest and chromatin accessibility. We refer to these cell types with known epigenetic signal and available chromatin accessibility as reference cell types. Because of the abundance of DNase-seq available in ENCODE [220], Epitome uses DNase-seq as the primary measure of chromatin accessibility for model training. However, Epitome can make predictions in a query cellular context using other assays that

measure chromatin accessibility, such as ATAC-seq (see Figure 2.7). We leverage chromatin accessibility from all reference cell types to compute an explicit metric of similarity between reference cell types and the query cellular context. This metric is referred to as the chromatin accessibility similarity vector (CASV), and compares the similarity between reference cell types and a query by comparing the similarity of chromatin accessibility at multiple windows of resolution surrounding a genomic locus of interest (See Methods 2.3.1.2). The CASV, along with binary epigenetic events from reference cell types, are used as features in a neural network, which uses the CASV to weigh the importance of each reference cell type for predicting epigenetic events in the query. Final predictions are the probabilities of observing each epigenetic event in question.

The set of candidate positions on which Epitome is applied consists of all loci in which the epigenetic event of interest has been observed in at least one of the reference cell types. This set of candidates is determined by unifying all the observed peaks from reference cell types. One important observation is that Epitome simplifies the input epigenetic data as binary, and the features it uses represent the presence or absence of "peaks" (or a signal). While this can lead to loss of information, it helps mitigate bias and noise arising from variations in sequencing depth, quality of antibodies and other technical factors that may affect the data quantitatively [185]. This strategy also facilitates the use of multiple types of epigenetic events as features (in addition to the mandatory accessibility data) without the need for calibration of their dynamic range or other quantitative attributes. Specifically, if a query cellular context has available measurements of certain histone modifications in addition to chromatin accessibility, these can be used to better evaluate similarity to the reference cell types and thus guide the prediction process (by extending the definition of CASV; see Figures 2.8(d) and 2.8(e)).

Underlying Epitome is a feed forward neural network (NN). Epitome's underlying NN can be written as $P(y|x, \mathbf{w})$. For the problem of predicting the presence of ChIP-seq peaks, $y \in [0, 1]$ is a set of classes, where each class is a ChIP-seq target predicted by Epitome. The values of y indicate the presence (1) or absence (0) of a ChIP-seq peak in a given 200bp region in the genome. $\mathbf{w}$ are a set of parameters learned by the Epitome model. x represents a set of features used to train an Epitome model. Features x contain ChIP-seq peaks from well characterized ENCODE cell types, as well as a measure of chromatin accessibility similarity between ENCODE cell types and query cellular context q. These features x are explained in greater detail in Sections 2.3.1.3 and 2.3.1.2. Figure 2.6 represents a schematic of Epitome. This figure visualizes how Epitome computes the similarity of chromatin accessibility between ENCODE cell types and q and uses this similarity as input into the model, along with binarized ChIP-seq peaks from ENCODE cell types, to predict peak probabilities for ChIP-seq targets in q. Model output are the probabilities of observing a peak for each ChIP-seq target being predicted in q at a 200bp genomic region i. Epitome can predict individual or multiple ChIP-seq targets in a single model, depending on which ChIP-seq targets and cell types are selected to train the model. We model the loss as the sigmoid cross entropy between the model's prediction of ChIP-seq targets $\hat{y}$ and the ground truth labels y. Cross

entropy loss is defined in defined in Equation 2.5.

$$loss = max(\hat{y}, 0) - \hat{y} * y + log(1 + e^{-|\hat{y}|}) \tag{2.5}$$

We parameterize the model, $P(y|x, \mathbf{w})$, by constructing cell type specific channels for each reference cell type used in training. Cell type specific channels are visualized in Figure 2.6 and explained in Section 2.3.1.3.

2.3.1.2 Measuring cell type similarity with the Chromatin Accessibility Similarity Vector (CASV)

As previously mentioned, Epitome uses a metric of cell type similarity to weigh the importance of reference cell types when predicting epigenetic signal in a query cellular context. This metric is referred to as the chromatin accessibility similarity vector (CASV). Figure 2.6 visualizes how the CASV is calculated between a query cellular context q and a reference cell type k at a 200bp genomic region i. The CASV compares the agreement of binarized accessibility peaks between q and k at varying genomic windows up to 12kbp surrounding i. Epitome uses the CASV to determine how similar q is to all reference cell types. This similarity is leveraged to determine the relative importance of each reference cell types in predicting ChIP-seq peaks for q. Without the CASV, Epitome is not able to provide cell type specific predictions for q, as shown in Figure 2.8(a).

We formally define the CASV in Equation 2.9. a^k and a^q indicate binarized chromatin accessibility peaks for cell types k and q binned in 200bp bins across the genome, where $a_i^k \in [0, 1]$ represents the presence (1) or absence (0) of a peak in bin i. The CASV calculates the fraction that a^k and a^q have shared chromatin accessibility peaks in region i. It also calculates the fraction that a^k and a^q have shared chromatin accessibility peaks in larger genomic windows surrounding region i. We consider exclusive 200bp genomic windows surrounding i in windows $R = \{r_z; 0 \leq z \leq 3\}$. We set $|r_0| = 1$, $|r_1| = 5$, $|r_2| = 19$, and $|r_3| = 59$, representing the number of 200bp bins considered in each window surrounding i. We first compute $CASV_n$, a vector of fractions of how many bins a^k and a^q agree for each window $r_z \in R$ surrounding region i, relative to the size of each window. Here, agreement criteria is met if both q and k have a peak or do not have a peak. Each window r is computed exclusively from smaller windows to avoid redundancy:

$$CASV_n(i, a^q, a^k) = \Big[\frac{\mathbb{1}[a_i^k = a_i^q]}{r_0}, \frac{\sum_{g \in r_1, g \notin r_0} \mathbb{1}[a_g^k = a_g^q]}{|r_1| - |r_0|}, \tag{2.6}$$

$$\frac{\sum_{g \in r_2, g \notin r_1} \mathbb{1}[a_g^k = a_g^q]}{|r_2| - |r_1|}, \frac{\sum_{g \in r_3, g \notin r_2} \mathbb{1}[a_g^k = a_g^q]}{|r_3| - |r_2|}\Big] \tag{2.7}$$

Because chromatin accessibility peaks are sparse across the genome for most cell types, regions where both q and k have a chromatin accessibility peak are rare. Therefore, high values of $CASV_n$ mostly represent regions of the genome where a^q and a^k do not have peaks.

However, to gain a complete understanding of cell type similarity, we must also know where a^q and a^k are both accessible. We encapsulate shared accessibility in $CASV_p$. This metric indicates shared regulatory regions between cell types [29].

$$CASV_p(i, a^q, a^k) = \Big[\frac{a_i^k \cdot a_i^q}{|r_0|}, \frac{\sum_{g \in r_1, g \notin r_0} a_g^k \cdot a_g^q}{|r_1| - |r_0|}, \frac{\sum_{g \in r_2, g \notin r_1} a_g^k \cdot a_g^q}{|r_1| - |r_1|}, \frac{\sum_{g \in r_3, g \notin r_2} a_g^k \cdot a_g^q}{|r_3| - |r_2|}\Big] \quad (2.8)$$

The final $CASV$ for a region i is a concatenation of the agreement and positive $CASV$ vectors:

$$CASV(i, a^q, a^k) = [CASV_p(i, a^q, a^k)||CASV_n(i, a^q, a^k)] \quad (2.9)$$

$CASV(i, a^q, a^k)$ are used as features in the Epitome model, described in Equation 2.10. The final $CASV$ comparing similarity between two cell types is a vector of length 8. Using the CASV, Epitome is thus able to learn different weights corresponding to elements in the CASV, and can thus determine the relative importance of chromatin accessibility similarity at varying distances from the genomic region of interest for each reference cell type.

In practice, Epitome uses binarized peaks called from DNase-seq to calculate the CASV during model training. However, ATAC-seq can also be used to compute the CASV. In the case of model training, the CASV measures similarity of binarized DNase-seq peaks between a held out cell type k' and each training cell type $k \in C, k \neq k'$, where C is the set of reference cell types in which the epigenetic event of interest has been measured. In the case of evaluating a query cellular context q, the CASV measures either binarized DNase-seq or ATAC-seq peak similarity between q and each reference cell type $k \in C$.

2.3.1.3 Constructing features from ENCODE cell types

Epitome trains on multiple reference cell types in a single model. These reference cell types are primarily taken from ENCODE. This allows Epitome to jointly learn from multiple cell types, without biasing models to overfit to a single training cell type. Therefore, features x for training a model contain cell type specific features for each cell type $k \in C$ that is selected for training. Here, we notate the number of cell types used to train a model as n, which is equivalent to $|C|$. Epitome can train on 2 to 93 ENCODE cell types in a single model. Because each reference cell type is uniquely characterized with different ChIP-seq targets, not all cell types can be utilized by Epitome to predict binding for a ChIP-seq target of interest. Therefore, the cell types chosen to train a model is determined by the number of cell types that have available ChIP-seq experiments for the ChIP-seq targets of interest. Figure 2.6 demonstrates an example schematic of an Epitome model in which three reference cell types are used for training a model that predicts an epigenetic event for a single ChIP-seq target.

The set of cell type specific features x^k for a cell type k selected for training is written in Equation 2.10. In Equation 2.10, F_i^k represents binarized ChIP-seq peaks in a 200bp region

i for the set of all ChIP-seq experiments that are available from cell type k. $CASV(i, a^q, a^k)$ is the chromatin accessibility similarity vector (CASV, see Section 2.3.1.2), and measures the similarity between chromatin accessibility in the query cell type, a^q, and the chromatin accessibility in the training cell type, a^k, at region i. $||$ represents concatenation.

$$x_i^k = [F_i^k || CASV(i, a^q, a^k)] \tag{2.10}$$

Finally, x_i^k is input into a cell type specific channel with two densely-connected layers. The input dimension for each cell type channel is the number of ChIP-seq targets included in the model and available in k, plus the dimension of the CASV(See Equation 2.9). This dimension can range from 9 (1 ChIP-seq target + 8 dimensional CASV) to 258 (the maximum number of ChIP-seq targets in a dataset + 8 dimensional CASV). The dimension of the first layer is the input dimension divided by 2. The dimension of the second layer is the input dimension divided by 4. Each layer uses a hyperbolic tangent activation function. The output of the last layer from each cell type channel is combined into a final output layer that applies a sigmoid non-linearity to gather final predictions $\hat{y}$.

2.3.1.4 Training an Epitome model

Epitome trains a model on n ENCODE cell types by using $n-1$ cell types for features x (defined in Equation 2.10) and a held out cell type as labels y. Epitome can train using up to 93 cell types in a single model, depending on the data availability of cell types for the ChIP-seq target being evaluated. Training on multiple cell types allows Epitome to generalize well to new cell types, and allows Epitome to perform comparably to methods that see the cell type being evaluated during training. To make full use of publicly available ChIP-seq experiments and effectively generalize to new cell types, Epitome uses a cell type rotation mechanism for training. This mechanism rotates through which cell type is used for y, and uses the remaining cell types to construct x. This rotation mechanism is repeated for all regions of the genome, excluding validation and testing regions, until the method converges. Algorithm 2 demonstrates how Epitome iterates through all training regions and all ENCODE cell types to update model parameters, using a different training cell type for y in each iteration. We determine convergence of Epitome by defining an early stopping criterion described in Morrow et al. [144] that determines when to stop training. This criteria halts training of models when validation loss no longer sees improvement.

Note that when a given cell type is used for labels y during training, its cell type specific features are still included in x through its respective cell type specific channel. Although this allows the model to see the labels during training, it also allows the model to learn the importance of cell type similarity in predicting ChIP-seq peaks. In this case, the model will compute identity cell type similarity between the cell type used for y and its features in x. This high similarity ultimately teaches the model to up-weight features from cell types that are similar to the cell type being predicted.

Algorithm 2 Epitome Rotation through cell types for model training

Require: C (set of training cell types)
Require: R (genomic regions selected for training)

```
for i in R do
    for k in C do
        x_i ← x_i^j ∀j ∈ C, j   ▷ x_i^j are cell type specific features as defined in Equation 2.10
        y_i^k ← ChIP-seq peaks for cell type k in region i
        loss = calculateLoss(x_i, y_i^k)
        updateGradients(loss)

        if model has converged then
            exit
        end if
    end for
end for
```

2.3.1.5 Sampling underrepresented ChIP-seq targets

When training a model for multi-label classification, imbalance across labels in the training data can result in a model that is biased towards learning over-represented labels. This problem is known as imbalanced learning. Imbalanced learning is present in models that predict peaks for multiple ChIP-seq targets because the quantity of peaks across the genome for different ChIP-seq targets is highly variable. As a solution to the problem of imbalanced learning when predicting ChIP-seq peaks, we over-sample underrepresented ChIP-seq targets to create a dataset with similar distributions of peaks for each ChIP-seq target in a multi-label model. This allows the model to see similar counts of positive instances for each ChIP-seq target during training.

To effectively over-sample underrepresented ChIP-seq targets, we borrow key insights from existing literature [28] which calculates an imbalance ratio, $IRLbl$, for each label to determine which labels need to be over-sampled. $IRLbl$ measures how imbalanced a given label in a dataset is, relative to the other labels. Higher values of $IRLbl$ indicate that a given label is more imbalanced, and thus needs to be over-sampled. $IRLbl$ is defined in Equation 2.11. In this equation, $D = (x_i, y_i)$ represents the multi-label dataset, L represents the set of labels (or ChIP-seq targets), and y_i represents labels for the ith instance of dataset D.

$$IRLbl(l) = \frac{\text{argmax}_{l'=L_1}^{L_{|L|}} \left(\sum_{i=1}^{|D|} h(l', y_i) \right)}{\sum_{i=1}^{|D|} h(l, y_i)}, h(l, y_i) = \begin{cases} 1 & l \in y_i \\ 0 & l \notin y_i \end{cases} \tag{2.11}$$

For each label l, $IRLbl(l)$ can then be compared to the mean imbalance ratio, $meanIR$,

for all labels $l \in L$. For each label l, if $IRLbl(l)$ is greater than the $meanIR$, we re-sample k_l regions in the genome that have a ChIP-seq peak for l. Here, we set k_l to be:

$$k_l = 10\frac{IRLbl(l)}{meanIR} * \sum_{i=1}^{|D|} h(l, y_i) \tag{2.12}$$

The final set of training instances includes all original instances in D, as well as instances oversampled for all labels $l \in L$ with $IRLbl(l) > meanIR$.

In the case where Epitome is only trained on one ChIP-seq target, multi-label sampling is not required. In this case, we undersample non-peak instances for the single ChIP-seq target so that there are 10 times more non-peak instances, compared to the number of positive peak instances.

2.3.1.6 Epitome achieves state-of-the-art accuracy for prediction of TFBS

To evaluate Epitome, we compared to four state-of-the-art methods that predict transcription factor binding sites, where each method is a best-in-class representative of methods designed to predict TFBS. For each TF and each method, we evaluated the ability to predict the binding landscape in a held-out cellular context given information from other cellular contexts that were used for training. The first benchmark method, DeFCoM [166] represents the class of footprinting methods, which use enzymatic cleavage patterns of DNase 1 or Tn5 transposase as features to predict TFBS. Because DeFCoM and other footprinting methods are motif centric, they are only designed to predict binding in regions centered around motifs. We therefore compared Epitome and DeFCoM by evaluating the ability of each method to predict TFBS overlapping motifs for 77 TFs and chromatin modifiers across 40 cell lines, primary cells, and tissues. Figure 2.9(b) shows scatter plots of performance for both Epitome and DeFCoM in regions overlapping motifs, where we compare the area under the precision recall curve (auPRC) and partial area under the receiver operating characteristic curve (pAUC) (5% FPR). We note that while the former measure accounts for all candidate binding sites, the latter measure is meant to highlight the accuracy of top predictions [128]. On average, both methods perform comparably under both metrics, even though Epitome is not isolated to training in motif regions. Additionally, out of the 280 comparisons, Epitome ranks number one for 202 and 199 experiments for auPRC and pAUC, respectively, while DeFCoM ranks number one for only 78 and 76 experiments for auPRC and pAUC, respectively (there were 5 ties for pAUC between methods). These results suggest that Epitome can perform better than motif centric footprinting methods that use explicit knowledge of both motif location and cleavage patterns to predict TFBS.

Although many footprinting methods are constrained to only predict in regions overlapping motifs, TFs can bind in regions not overlapping canonical motifs associated with the TF of interest [44]. This is demonstrated in Figure 2.2, which shows that the ratio of ChIP-seq peaks that do not overlap any motif for a TF in question can range from 0.45 to 1.0 for the 77 TFs and chromatin modifiers we evaluated. Because limiting the prediction regime to

sites that overlap motifs can eliminate many TFBS, Epitome and other methods have been designed to predict epigenetic events genome wide to eliminate this constraint. One example of such methods is Catchitt [100], a co-winner of the ENCODE-DREAM TFBS prediction challenge [108], which uses a combination of features including DNA sequence content, motif hits, expression levels of transcription factors (from RNA-seq), and chromatin accessibility (DNase-seq) to predict TFBS. Similar to DeFCoM, Catchitt considers accessibility data in a quantitative form, albeit at a default binned resolution of 50bp. It also uses a wider genomic context, taking up to 2,000 bp around the candidate site of interest. Additionally, Avocado [186] represents a class of imputation methods, designed to impute missing signal for cell types and assays that have not been measured experimentally. To compare to these methods, we first evaluated chromosome wide predictions of Epitome, Avocado, and Catchitt, on all 77 TFs and chromatin modifiers across 40 cell lines, primary cells, and tissues on held out chromosomes 8 and 9 in Figure 2.9(a), for a total of 264 comparisons. Here, we consider two versions of Epitome: single, which trains an individual model for each TF, and joint, which trains a single model to predict all TFs. These models are separately considered because Catchitt uses a single prediction approach, while Avocado uses a joint approach. Figure 2.9(a) shows that of the four methods, Epitome models trained individually for each TF perform the best for a majority of the 264 experiments. While joint Epitome models perform similarly in terms of pAUC and better for many instances in terms of auPRC compared to single Catchitt models (Figure 2.3(a)), Avocado performs poorest on these metrics (Figure 2.9(a)). Figure 2.3(b) additionally shows that for many cases, single Epitome models perform better than jointly trained Epitome models. It is unsurprising to note that single models often perform better than joint models. In these cases, learning objectives for different TFs can have complex or competing dynamics [200]. As shown in Figure 2.4, different ChIP-seq targets trained in a joint model converge at different iterations during training. This difference in convergence across targets can result in variance in the level of fit for each ChIP-seq target. Regardless, we find that overall performance of both joint and single Epitome models is higher than the other two methods in the majority of the 264 experiments, based on both auPRC and pAUC metrics (Figure 2.9(a)).

We also consider a different group of methods that use NN architectures to predict TFBS from DNA sequence, but are not able predict on new cell types. This includes DeepSEA [240], which uses a convolutional neural network to learn the mapping from DNA sequence to TFBS. Surprisingly, we find that the accuracy of Epitome compares similarly or better to these types of models, which use all the data during training (i.e., not leaving out one of the cell line as query). This class of methods use DNA sequence alone as features to predict binding of transcription factors. While the goal of DeepSEA is to predict the effect of sequence mutations on the epigenetic events it is trained with (rather than predicting events in unobserved cellular contexts), the fact that it sees all data during training provides a conceptual upper bound for accuracy. To compare to DNA sequence based methods, we compared the performance of a set of 17 TFs which were assayed by the ENCODE consortium in four cell lines that were available in DeepSEA models, resulting in 68 comparisons. All data used is from ENCODE, and was aligned to the hg19 genome. Figure 2.10 shows that

even though DeepSEA trains using a given test cell line and Epitome does not, Epitome performs comparably to DeepSEA in both pAUC and auPRC. These results demonstrate that Epitome can effectively generalize to unseen cellular contexts.

2.3.1.7 Epitome places an upper bound on maximum achievable sensitivity

Because Epitome constructs features using binarized epigenetic events in reference cell types, models are a-priori limited to predict epigenetic events in regions in which the epigenetic event in question has been observed in a previous experiment. Although this limitation could be alleviated by using genome-wide continuous signal, which is available in all genomic regions, we explain in Section 2.3.1.1 our explicit choice to use binary epigenetic signal to reduce noise. In Figure 2.5, we calculated the fraction of unique peaks observed a held out cell type for all ChIP-seq targets available in ChIP-Atlas [158] to determine an upper bound for the sensitivity that could be achieved using Epitome, given this limitation. When more than 26 cell types are available for a given ChIP-seq target, we observe that sensitivity exceeds 90%, on average. Furthermore, we observe that the upper bound on sensitivity increases quickly as the number of available reference cell types increases. While this approach places a strict upper bound on the achievable sensitivity, in Figure 2.9 we demonstrate that this strategy leads to a better overall balance between sensitivity and specificity, compared to existing methods.

2.3.1.8 Models trained on multiple cell types similar to the query cell type provide improved accuracy

Of the methods compared to in Section 2.3.1.6 that provide predictions of TFBS specific to a cellular context of interest, none are designed to jointly train on multiple reference data sets. This limitation bypasses the opportunity to jointly learn from multiple cell types and can generate misguided predictions when the query context greatly differs from the cell types selected for training. As previously mentioned, Figure 2.5 shows that as the number of cell types available for a given epigenetic event measured with ChIP-seq increases, the fraction of unique TF binding sites or histone modifications observed in a new cell type decreases. From this trend, we conclude that as more cell types for a given event are used to train a model, it is more likely that the model will have seen that event in the genomic location we are trying to predict. We therefore sought to understand the effect of the quantity and choice of reference cell types used for training Epitome on accuracy when predicting epigenetic signal in a new cellular context.

To assess the effect of the number of cell types used during training on model performance, we trained models for 82 different ChIP-seq targets, which included TFs, histones, and histone modifications. For each ChIP-seq target, we considered a range of models trained on different numbers of ENCODE cell types as references (ranging from 2 to $n-1$ reference cell types, where n is the number of cell types available for the respective ChIP-seq target). Cell types used include cell lines, in-vitro differentiated cells, primary cells, and tissues from

ENCODE [141]. For each ChIP-seq target and choice of cell type count used for training, we trained four models with different combinations of training and validation cell types. Each of these models were evaluated by predicting peaks across validation chromosome 7, which was held out from training. This procedure resulted in training $4 * (n - 2)$ models for each of the 82 ChIP-seq targets, where each model trained on a different number and combination of reference cell types.

Figures 2.11(a) and 2.11(b) demonstrate the change in performance (auPRC) in predicting TFBS with increasing numbers of reference data sets. Across the 59 TFs included in this analysis, we observe an overall consistent increase in performance as one considers larger numbers of reference cell types in which the TF in question was measured. While for most TFs, the number of cell types for which information was available is limited (under thirteen), the ENCODE collection includes over forty contexts for the CCCTC-Binding Factor CTCF. Considering the performance of Epitome in predicting the binding positions of CTCF, we observe that the value of adding more reference data sets starts to diminish at approximately ten data sets, and that beyond this point the performance tends to saturate. Since a similar level of saturation is not reached in the other TFs, these results may serve to provide intuition for the number of cell types which may be required to achieve high accuracy in future applications of Epitome.

Interestingly, when Epitome is applied to predict the positions of histones or histone modifications, we observe a similar trend of improvement in performance, but only up to a certain level. Similar to the TFBS prediction task, the performance of Epitome starts to saturate when more than ten data sets are used as a reference. However, we also observe a marked decrease in performance when the number of cell types that are used as reference goes beyond twenty. Taken together, these results suggest an optimal regime for the number of data sets to be used by Epitome as a reference. We note that this actual number can be evaluated by cross-validation, a utility which will become crucial as the number and diversity of ChIP-seq data sets increases.

2.3.1.9 Considering wide genomic contexts and multiple epigenetic signals to compute cell type similarity improves the performance of Epitome

Epitome uses the chromatin accessibility similarity vector (CASV) to estimate the extent to which a query cellular context is similar to the reference cell types at each candidate locus. The CASV accounts for similarities at several scales, from a small window of 200bp around the locus in question, up to a window of size 12kbp. We next explored whether accounting for multiple levels of resolution aids in learning better decision rules. To this end, we evaluated the performance of Epitome in predicting each of the 82 epigenetic events, shown in Figure 2.11, while varying the size of the genomic context used for the CASV. We considered five possible sizes of genomic context to be considered by the CASV, including 0bp (the identity CASV), 200bp, 1,200bp, 4,000bp, and 12kbp. For each of the models trained in Figure 2.11, we trained 4 additional models, each using a different genomic context ranging from 0-12kbp. These models were similarly trained using different numbers of reference data

sets ranging from 2 to 30 in size, based on the availability of data sets for a given ChIP-seq target.

Figure 2.8(b) and 2.8(c) depict the overall change in pAUC and auPRC, respectively, as larger genomic contexts are considered. The most basic models, which simply tally up the number of observed peaks (i.e., do not account for accessibility data; "identity CASV") as a decision rule perform the poorest, thus supporting the merit of using chromatin accessibility data. Consistently, we observe slight increases in performance as the genomic context considered by the CASV increases. This observation is most easily observed in change in AUC as wider genomic contexts are considered, shown in Figure 2.8(a). We find that while performance increases with the presence of a wide genomic context for models that are trained with less than 10 cell types, histone modifications in particular suffer from considering wide genomic context when models are trained on more than 10 cell types. Figures 2.12(b) and 2.12(a) show that when histone modifications are trained on more than 10 cell types, performance decreases as wider genomic contexts are considered. These observations agree with trends seen in Figure 2.11(c), showing that performance of histone modifications does not saturate, but degrades, with increasing information in models trained with more than 10 cell types.

In all results shown thus far, Epitome only incorporates chromatin accessibility in the CASV to compute similarity between a query cellular context and reference cell types used for training. However, additional assays, such as ChIP-seq for histone modifications, are often available in addition to chromatin accessibility, and could be used to compute a more robust measure of similarity of a local chromatin environment. Histone modifications in particular provide valuable information to compute chromatin similarity, as it has been observed that DNase-seq hypersensitivity sites that are common to many cell types are only weakly correlated with certain histone marks [192]. In these cases, weakly correlated histone marks can provide somewhat independent and potentially more specific information of similarity in regions that already have similar chromatin accessibility. This is also a common use case that a query cellular context has been partially characterized with multiple experiments, including chromatin accessibility and various histone modifications (Figure 2.13).

To make use of all assays that may partially characterize a cellular context, we explored whether extending the CASV to include histone modifications in addition to chromatin accessibility could improve predictive accuracy of Epitome. To test this, we evaluated the effect of incorporating various histone modifications in the CASV, and how this alteration changed the auPRC performance for thirteen TFs, listed in Figure 2.8(d). These thirteen TFs were selected in particular because they were available in the four reference cell lines that had measurements for seven histone modifications, allowing us to extend the CASV to use histone modifications from reference cell types. The seven histone modifications incorporated into the CASV included H3K9ac, H3K4me3, H3K4me2, H3K4me1, H3K36me3, H3K27me3, and H3K27ac.

Incorporating these seven histone modifications, as well as DNase-seq, in the CASV resulted in 256 configurations of Epitome that used all possible combinations of the seven histone modification and DNase-seq to compute the CASV. For each configuration, we trained

four separate models using three reference cell lines each, and then evaluated each model on a fourth held out cell line. Final auPRC performance for each configuration was calculated across predictions for all four evaluated models.

Figure 2.8(d) shows the difference between auPRC performance of Epitome using a single histone modification as well as DNase-seq in the CASV and auPRC performance using only DNase-seq in the CASV. A majority of these thirteen TFs see minor improvement in auPRC performance when including a histone modification, with the exception of YY1, TCF12, and TBP. All histone modifications evaluated have some positive improvement in auPRC, with the exception of H3K9me3. This observation is consistent with the association of H3K9me3 with the formation of transcriptionally silent heterochromatin.

We next sought to understand the relative contributions of DNase-seq and the seven histone modifications in the CASV towards model performance for the 13 TFs evaluated in Figure 2.8(d). We computed the Shapley value for each combination of experiments used in the CASV for each of the 13 TFs. The Shapley value indicates the marginal contribution computed from all possible subsets [73]. These values ultimately can highlight which experiments provide the maximal information for predicting TFBS, and which experiments should be prioritized when partially characterizing a cellular context of interest. Shapley values are shown in Figure 2.8(e) for each histone modification and DNase-seq, and their effect on each of the 13 TFs evaluated. On average, repressive mark H3K9me3 gives the least information when incorporated into the CASV, and in most cases, provides negative contribution to auPRC performance. We hypothesized that H3K9me3 gave the least amount of information because its correlation with TFs was low, providing the model with little to no consensus information. To test this hypothesis, we computed the Jaccard index between TFBS in each 200bp window for each TF in each training cell line and each peak indicating a histone modification used in the CASV. Indeed, H3K9me3 had the smallest mean Jaccard score across training cell lines (0.0008), compared to other histone modifications and DNase-seq (0.04 mean). These results suggest that including histone modifications in the CASV can improve predictive performance when it correlates with the epigenetic signal of interest.

2.3.1.10 Epitome recapitulates changes in H3K27ac over neural induction of human pluripotent stem cells

We next evaluated Epitome's ability to leverage changes in ATAC-seq to detect changes in the acetylation of H3K27 (H3K27ac) over neural induction of human pluripotent stem cells (hPSCs). This analysis can demonstrate how well Epitome is able to leverage changes in chromatin accessibility from the same starting population to detect gradually accumulating H3K27ac marks. Temporal analysis of neural induction from hPSCs has shown that changes in chromatin accessibility precede H3K27ac, a histone mark indicative of transcriptionally active regions [83]. We therefore sought to use Epitome to predict H3K27ac in seven timepoints of neural induction by using chromatin accessibility to compare timepoints to reference cell types used to train a model.

In previous analyses of neural induction, Inoue et al. collected H3K27ac and ATAC-seq across seven time points of neural induction starting from hPSCs. Genomic regions that were enriched for H3K27ac over timepoints were grouped into six clusters, each associated with a different temporal pattern. These clusters identified H3K27ac peaks present in early induction (clusters 1-3), mid induction (cluster 4), and late induction (clusters 5-6). To determine whether Epitome could leverage ATAC-seq to identify changes in H3K27ac over neural induction, we used ATAC-seq from each time point as input into the CASV to predict H3K27ac peaks. Although the Epitome model primarily uses DNase-seq to compute the CASV and train its models, we show in Figure 2.7 that Epitome can accurately predict histone modifications and TFBS when using ATAC-seq during evaluation in the CASV when Epitome has been trained using DNase-seq. For this reason, we hypothesized that Epitome could provide sensitive predictions of H3K27ac by using the CASV to compare similarity between DNase-seq in reference cell types and ATAC-seq from neural differentiation time points. We trained an Epitome model to predict H3K27ac peaks using all ENCODE reference cell types that had both H3K27ac and DNase-seq, resulting in 15 reference cell types.

Here, we compare Epitome to an additional benchmark method specifically designed to predict histone modifications, called DeepHistone [233]. DeepHistone is a deep learning method for predicting seven histone modifications, including H3K27ac, and uses DNA sequence and chromatin accessibility as features to provide predictions specific to a cellular context. For further insight we also added a simple baseline predictor, which uses enrichment signal from ATAC-seq peaks to directly indicate the signal of H3K27ac (i.e., predict an H3K27ac peak whenever there is an ATAC-seq peak). We chose this baseline because increased acetylation of H3K27 is often observed in accessible regions, and in neural induction it was often observed to be preceded by opening of the chromatin [83]. Figure 2.14(a) shows the ROC and PR curves for H3K27ac peaks across seven time points in 39,000 regions for the three comparative methods. All autosomal regions that had an H3K27ac peak in at least one of the seven timepoints were considered. For all time points, Epitome performs significantly better than the baseline predictor as well as DeepHistone, with a gain of mean auPRC of 0.09 and 0.21, over the baseline ATAC-seq predictor and DeepHistone, respectively. This boost in performance seen by Epitome suggests that using known H3K27ac peaks from ENCODE cell types, along with the CASV to inform the model of similarity between ENCODE cell types and differentiated hPSCs, improves predictive performance of H3K27ac predictions over baseline predictors that do not use signal from the epigenetic event of interest as features.

Although Epitome can predict H3K27ac peaks better than baseline predictors and existing models, ROC and PR curves could not illuminate whether Epitome could identify H3K27ac peaks that were unique to each time points. We therefore sought to show that Epitome can leverage changes in chromatin accessibility over time to detect time point specific H3K27ac peaks. Figure 2.14(b) demonstrates the mean predictions of Epitome (top) and the normalized H3K27ac read counts (bottom) in six H3K27ac clusters indicating key H3K27ac peaks of early, mid, and late induction across 2,400 genomic regions. These six clusters were defined by Inoue et al. [83]. Epitome predictions show that at 48hr and 72hr,

there is an increase in H3K27ac in clusters 5 and 6. Additionally, clusters 2, 3, and 4 have increased H3K27ac between time points 3hr to 24hr, and decrease at 72hr. These broad trends of increased H3K27ac in late induction clusters 5 and 6 and in early to mid induction clusters 2-4 agree with H3K27ac normalized read counts shown in the bottom heatmap.

In this analysis, Epitome leverages ATAC-seq to predict time point specific H3K27ac. This means that for a given region of the genome, Epitome will predict the same probabilities for H3K27ac for all samples, unless ATAC-seq signal varies across samples in or around a given region of interest. Because of this, we would expect Epitome to provide sensitive results when predicting H3K27ac in regions with differential ATAC-seq between time points. To assess this hypothesis, we evaluated Epitome H3K27ac predictions at 0hr and 72hr, and separated predicted regions into three groups: regions accessible in all time points, regions only accessible in either 0hr or 72hr, and regions not accessible in either 0hr or 72hr. Figure 2.14(c) shows the CDFs of Epitome H3K27ac predictions at 0hr and 72hr across these three groups. At both the 0hr and 72hr time points, Epitome predictions are higher in accessible regions, compared to inaccessible regions. These results suggest that Epitome is effectively leveraging accessibility to predict H3K27ac. We note that for both the 0hr and 72hr predictions, Epitome is able to easily detect H3K27ac in shared accessible regions of the genome, compared to regions of the genome that have differential accessibility in either the 0hr or 72hr time points. In regions of differential accessibility, Epitome predictions are lower than those found in regions of shared chromatin accessibility. This is most likely an artifact of decreased availability of evidence for H3K27ac in the training cell types in regions that have differential accessibility. In regions of shared accessibility, training cell types have H3K27ac marks in a median of 8 out of 13 cell types, compared to peak regions unique to either 0hr or 72hr time points, which have H3K27ac marks in a median of 2 out of 13 cell types. Thus, commonly accessible regions have greater evidence of H3K27ac marks in training cell types, and are predicted with greater confidence.

Although Epitome can predict H3K27ac marks better than existing methods, Epitome still incorrectly identifies a subset of peaks as false negatives, and a subset of non-peak regions as false positives. Figure 2.14(d) and Figure 2.15(a) show Epitome predictions for H3K27ac at 72hr in peak and non-peak regions respectively. Figure 2.14(d) demonstrates that while a majority of H3K27ac marks are correctly identified, there is a small subset of peaks that are incorrectly identified as false negatives. Additionally, many of the non-peak regions are identified as peaks. Because of this trend, we sought to identify the source of the false positives and negatives within the context of the reference cell types used for training. Figure 2.14(d) shows reference data sets used as input in the Epitome model for predicting H3K27ac marks in peak regions at 72hr. Epitome's true positives are shown towards the bottom of the heatmap, while false negatives are displayed at the top. This plot shows that true positives have ATAC-seq accessibility at the 72hr time point, as well as supporting H3K27ac peaks from the reference cell types used for training. False negatives generally have less ATAC-seq accessibility at 72hr, and less H3K27ac peaks from the reference cell types. These results demonstrate that false negatives were generated from regions that had little to no accessibility at the 72hr timepoint, as well as little support for H3K27ac in reference cell

types. This trend is similarly shown in the predictions of non-peak regions at 72hr, shown in Figure 2.15(b), where false positives, shown towards the bottom of the heatmap, have more accessibility at 72hr and representation of H3K27ac peaks in the reference cell types than peaks correctly identified as true negatives. These results show that false positives and false negatives result from unexpected patterns in chromatin accessibility and H3K27ac in both the query cellular context and reference data sets.

2.3.2 scEpitome: Predicting epigenetic events, one cell at a time

As discussed in Section 2.3.1, Epitome is a method that can leverage chromatin accessibility from DNase-seq or ATAC-seq to predict epigenetic events in novel cellular contexts. However, DNase-seq and ATAC-seq only provide us with an "average" of chromatin accessibility, as each sample is prepared and sequenced from hundreds of thousands of cells. Recent advances, such as single-cell assay for transposase-accessible chromatin (scATAC-seq), are able to capture variation in chromatin accessibility at single cell resolution. Similar to ATAC-seq, scATAC-seq uses Tn5 transposase to tag regulatory regions of the genome. However, unlike ATAC-seq, scATAC-seq uses cell-identifying barcoded primers to track and identify variability of accessibility between single cells [26, 38]. scATAC-seq allows us to answer many questions that were previously impossible using bulk measurements of chromatin accessibility, including the ability to characterize heterogeneity of chromatin accessibility in tumor samples [111], identify markers of sub populations of cell types within larger samples [179], and determine the effect of knocking out components crucial to biological pathways on variability of accessibility within a sample [47]. In this section, we discuss an extension of Epitome (Section 2.3.1), called scEpitome, that predicts TFBS at single cell resolution by leveraging scATAC-seq. We compare scEpitome to state-of-the-art methods that predict changes in regulatory activity at single cell resolution, and demonstrate that Epitome achieves better sensitivity and precision when using bulk ChIP-seq data as ground truth. This section consists of unpublished work.

In the advent of scATAC-seq, numerous methods have been developed to interpret the regulatory effect of changes in chromatin accessibility at single cell resolution. One such example is chromVAR, which was primarily designed to estimate deviations in the accessibility of TF motifs between single cells. chromVAR estimates changes in accessibility of peak regions called from scATAC-seq that overlap motifs of interest [181]. chromVAR corrects for GC content and mean accessibility within a dataset by correcting for background motif overlap to provide bias corrected deviation scores for each motif and cell of interest. However, chromVAR primarily uses motif information to approximate the activity of TFBS at single cell resolution. As discussed in Section 2.3.1 and demonstrated in Figure 2.2, TFs can bind in regions not overlapping canonical motifs [44]. This implies that methods using motifs as an indicator of TFBS may be overlooking binding sites with non-canonical motifs.

To improve on existing approaches and eliminate drawbacks of motif-centric approaches that annotate changes in chromatin accessibility from scATAC-seq datasets, we propose a modification of Epitome [144] that supports prediction of TFBS by leveraging scATAC-seq

datasets. As input, scEpitome takes in adjusted read counts overlapping each called peak region for each single cell. Ideally, these read counts would be taken directly from aligned reads from the scATAC-seq dataset of interest. However, various factors, such as low sensitivity, batch effects, and variation in sequencing coverage, make direct counting of reads from scATAC-seq in downstream analysis pipelines challenging. Therefore, we leverage PeakVI [8], which estimates the probability of each peak being accessible in each cell, and takes into account technical factors that would affect the probability of accessibility. Probabilities from PeakVI for each peak and cell are used as input into scEpitome. Ideally, we would compare single cell accessibility to a reference set of single cells and incorporate this information into the CASV 2.3.1.2 to gather measures of similarity between cells in our dataset and cells in the reference. However, there is currently no reference scATAC-seq datasets with matching experimental information of TF binding at single cell resolution. Therefore, similar to Epitome, we compare PeakVI probabilities of accessibility for each peak and cell in question to bulk DNase-seq data from well annotated reference cell types. This similarity is incorporated into the CASV (See 2.3.1.2) to determine the level of similarity between each single cell and each bulk reference cell type. Similar to Epitome, scEpitome uses the CASV to compute cell type similarity between accessibility of single cells and available reference cell types. However, unlike Epitome, which was designed for consumption of bulk ATAC-seq and DNase-seq, scEpitome uses probabilities of peak accessibility, instead of binary peak values. Therefore, we modify the CASV to work for continuous values of accessibility, rather than binary peaks. To do so, we modify Equation 2.9 by eliminating $CASV_N$ (Equation 2.7) from the CASV, as it assumes exact similarity between binary peaks. Thus, when leveraging scEpitome to predict TFBS from scATAC-seq, we only compare the probability of peaks in single cells to the presence of peaks in reference cell types (Equation 2.8).

To evaluate the ability of scEpitome to predict TFBS at single cell resolution, we evaluate a scATAC-seq dataset consisting of 63,882 peripheral blood mononuclear cells (PBMC) and bone marrow single cells published from Satpathy et. al. [179]. Figure 2.16(a) visualizes latent representations, produced from PeakVI, of all single cells. Cells are colored by population, as identified by magnetic-activated cell sorting (MACS) or FACS [179]. Posteriors for each peak and cell, generated from the latent representation in Figure 2.16(a), were used as input into Epitome as an indicator of accessibility at single-cell resolution.

Although single-cell chromatin immunoprecipitation followed by sequencing (scChIP-seq) has been previously measured [68], there are currently no matched scChIP-seq and scATAC-seq datasets available. Thus, we are unable to leverage scChIP-seq as ground-truth to evaluate prediction of TFBS in scATAC-seq datasets. We therefore used bulk ChIP-seq data from ChIP-Atlas [158] to verify predictions of TFBS in sorted populations. We predicted binding sites for 17 TFs that had available bulk ChIP-seq data in ChIP-Atlas in a subset of cell types identified by FACS. This included analysis of the following TFs: YY1, STAT3, SPI1, REST, RELA, NR3C1, NFKB1, NFE2, MEF2B, JUNB, FOXP1, FOSL1, FOS, ETS1, EGR1, CTCF, and CEBPB. For these TFs, ChIP-Atlas had bulk ChIP-seq data for six cell types identified through FACS, including: monocytes, dendritic cells, hematopoietic stem cells (HSC), B cells, PBMCs, CD34 progenitors, and naïve CD4(+) T cells. To gather

predictions of TFBS across these six cell types, we averaged posteriors of accessibility across all single cells annotated within the cell type in question. These averages gave us the expected accessibility for each peak in a given cell type. These expected values were used as input into Epitome to gather predictions of TFBS.

To evaluate scEpitome's ability to predict TFBS in single cells, we compared to two alternative approaches for prediction of TFBS. chromVAR directly computes motif scores for each single cell by computing a deviation score for each motif and cell of interest, instead of providing region-specific predictions for each motif. Therefore, we were unable to directly use chromVAR to predict TFBS at peak resolution. Instead, we leveraged chromVAR's method of correcting fragment counts with a background. Here, chromVAR calculates a background for each peak, based on GC content and average accessibility in a dataset. chromVAR calculates this background by sampling background peaks similar in GC content and average accessibility to each peak in the dataset, sampling from a normal distribution $\sim \mathcal{N}(0, 0.1)$. We use this sampling strategy to select background peaks for each peak in our dataset. We gather corrected fragment counts by subtracting fragments in each sampled background region from each peak in the dataset. To gather predictions of TFBS, for each population, we calculate the expected fragment counts for each peak in each of the six validation cell types. We then mask each region by the presence or absence of a motif for the TF of interest, giving us peak signal in all regions overlapping a motif. Motifs were taken from the JASPAR database [55].

We also compare to a baseline method that directly uses PeakVI posteriors to gather predictions of TFBS, instead of background corrected fragment counts, as used in chromVAR. To gather predictions of TFBS using this baseline method, we calculate the expected accessibility for all peaks in each of the six validation cell types, identical to the input of scEpitome. We then mask each region by the presence or absence of a motif for the TF of interest, giving us peak signal in all regions overlapping a motif.

We first compared our baseline method using PeakVI posteriors to chromVAR background corrected reads towards the prediction of TFBS in pseudo-bulk populations from scATAC-seq. Surprisingly, Figures 2.16(b) and 2.16(c) show that, when using bulk ChIP-seq as ground truth, our baseline method provides a better basis for analysis of motif overlaps than chromVAR background corrected fragments. Because of this, we next used the baseline method and compared it to scEpitome predictions. Figures 2.16(d) and 2.16(e) shows that scEpitome performs better than the baseline method, in terms of auPRC and pAUC, for a majority of the TFs and FACS-isolated populations evaluated. These results suggest that leveraging available ChIP-seq and chromatin accessibility similarity between bulk and single cell populations can be used to gather more accurate predictions of TFBS in small populations isolated in scATAC-seq datasets.

Although this PBMC and bone marrow dataset provides us with FACS-isolated populations, this information is most often not available for many scATAC-seq datasets of interest. Therefore, we wanted to use a computational method to identify clusters in large scATAC-seq datasets, and then leverage scEpitome to predict TF binding sites on smaller clusters. To identify smaller clusters, or microclusters, we used VISION [40], which merges single cells

with similar profiles into representative pools. We used VISION to identify microclusters with about 500 cells per cluster, resulting in 128 microclusters. Figure 2.17(a) shows UMAP representations from Figure 2.16(a) for all cells for each of the 128 microclusters identified by VISION.

We next predicted TF binding in each microcluster for the 17 TFs available in ChIP-Atlas and bone marrow derived populations. We calculated the expected accessibility from PeakVI posteriors for each peak and microcluster. Although no ground truth labels of TF binding was available for microclusters, we hypothesized that clusters that had larger fractions of cells that were labeled as B cells, dendritic cells, or one of the six cell types listed in Figure 2.16(b), would have more similarity of binding patterns to bulk corresponding ChIP-seq than microclusters that were less similar to the bulk cell type in question. Figure 2.17(b) shows the slope of the curve calculated between similarity of chromatin accessibility between each microcluster and bulk DNase-seq experiment to the similarity between bulk ChIP-seq and scEpitome predicted binding sites. We would expect this slope to be positive, as populations with similar chromatin accessibility should predict similar TF binding patterns. While TFs such as SPI1, STAT3, and FOXP1 have strong association between chromatin accessibility and binding similarity, others have weakly or slightly negatively associated slopes. These results suggest that while some TFs have variable TF binding predictions between microclusters, others have little change in similarity of predictions (Figure 2.17(c)).

In conclusion, we have shown that scEpitome can predict TFBS in microclusters calculated from scATAC-seq datasets, and that these predictions can suggest finer grained changes in patterns of TF binding across clusters than when leveraging bulk measurements of chromatin accessibility. We have shown that scEpitome can predict binding patterns of TFs more confidently than baseline methods that directly use scATAC-seq as a predictor for binding, as well as methods designed to learn general TF binding patterns across single cells, when using bulk ChIP-seq datasets as ground truth. These initial results suggest that scEpitome can be leveraged to identify regulators with differential activity between sub-populations within cell types that cannot be identified using bulk measurements of chromatin accessibility. Future work first requires scaling scEpitome to predict binding at single cell resolution, rather than at the level of sub-populations. Currently, scEpitome computes predictions of TFBS for each single cell sequentially. Ongoing work involves scaling scEpitome to predict TFBS at single cell resolution in parallel. This extension will support discovery of regulators that have differential binding patterns across single cells within sub-populations.

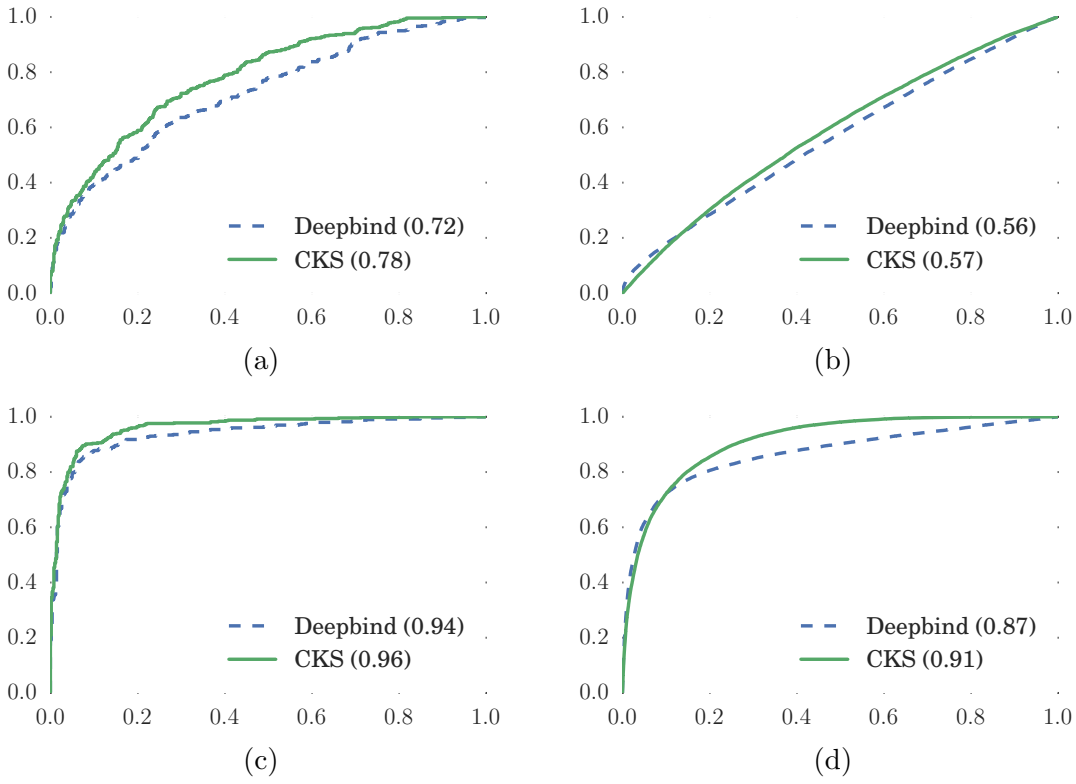


Figure 2.1: Receiver operating characteristic curve (ROC) of DeepBind and CKS for transcription factors EGR1 and ATF2 for GM12878. (a) ROC for ATF2 on DeepBind's test set. (b) ROC for ATF2 on ENCODE peaks. (c) ROC for EGR1 on DeepBind's test set. (d) ROC for EGR1 on ENCODE peaks.

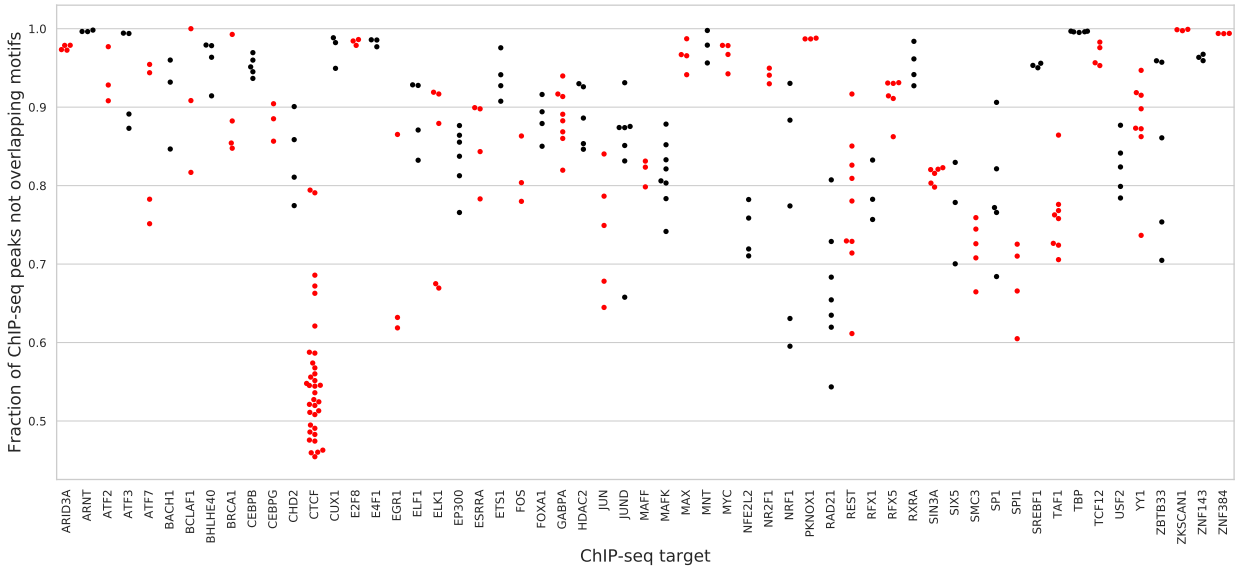


Figure 2.2: Ratios of ChIP-seq peaks from ENCODE (hg38) that do not overlap any motif. 77 TFs and chromatin modifiers were considered across 40 cell types. Each data point represents the ratio of motif misses for a ChIP-seq experiment from a TF/chromatin modifier and cell type combination.

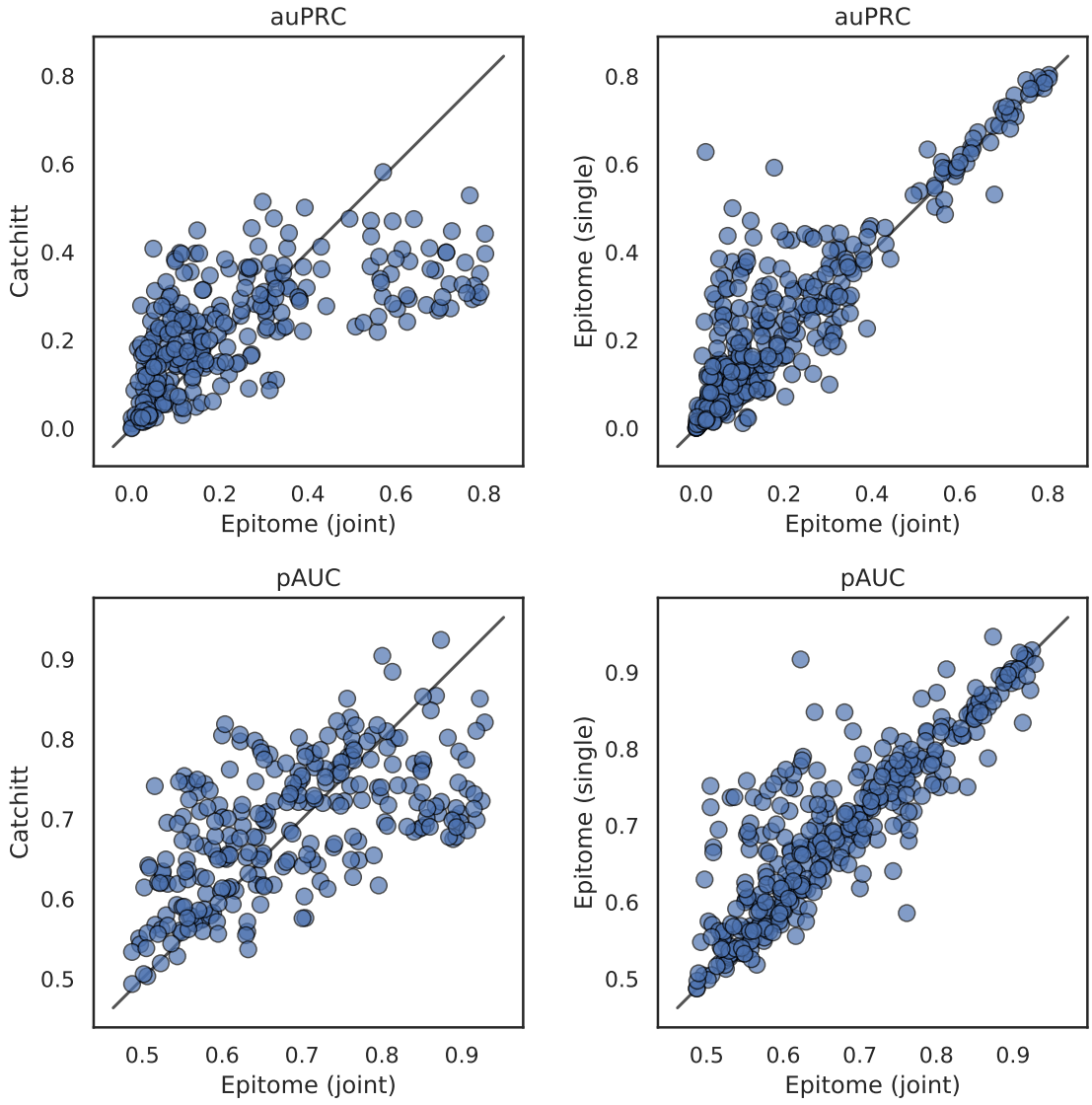


Figure 2.3: Performance of Epitome joint models, single models, and Catchitt for predicting ChIP-seq peaks for 77 transcription factors on chromosomes 8 and 9 in 40 held out primary cells, tissues, and cell lines. (a) auPRC and pAUC (5% FPR) scores for Epitome joint models and Catchitt. (b) auPRC and pAUC (5% FPR) scores for Epitome models trained jointly and Epitome models trained individually (single) for each TF.

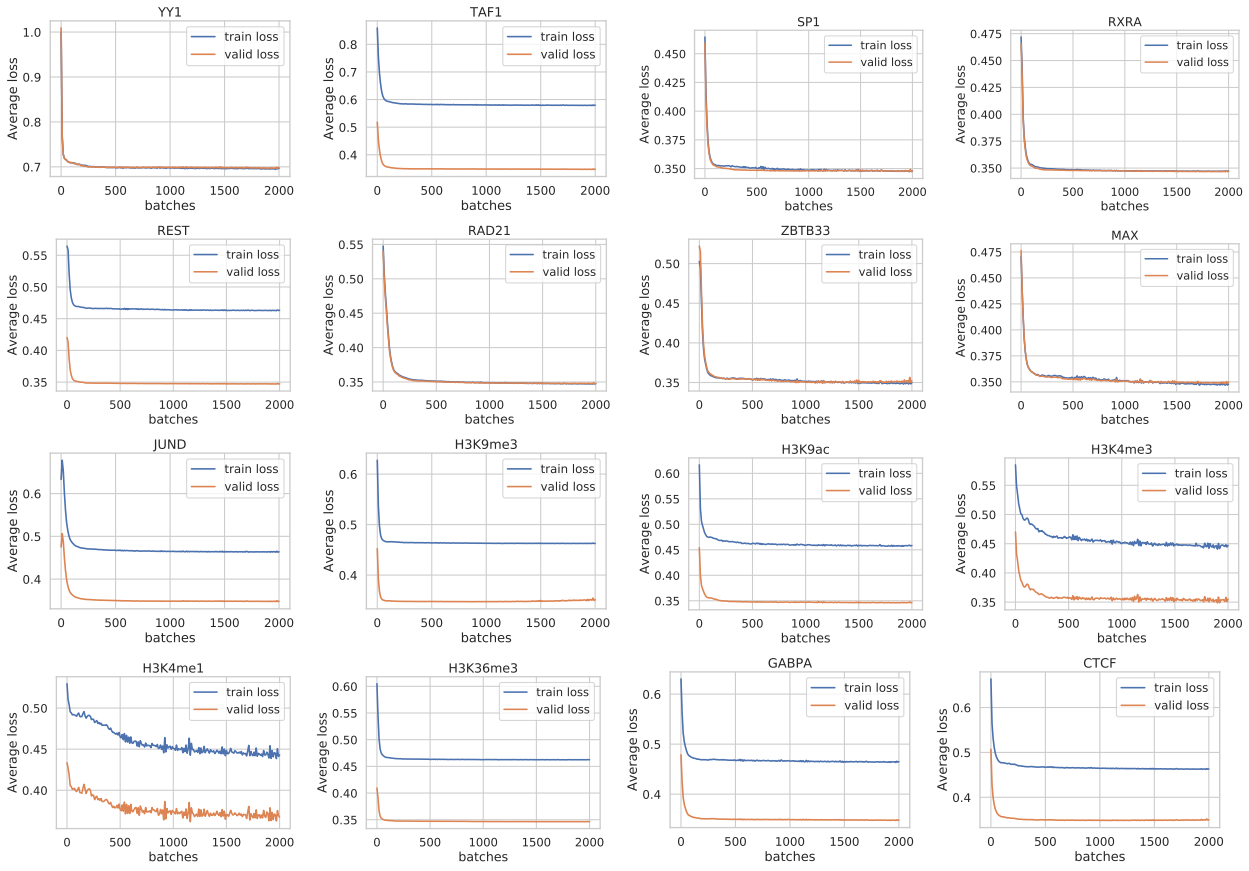


Figure 2.4: Average training and validation loss for 16 ChIP-seq targets, including transcription factors, chromatin modifiers and histone modifications. All ChIP-seq targets were trained jointly. Average train loss was calculated from 10,000 sampled points from the training dataset. Valid loss was calculated from all points on chromosome 7 meeting the sampling criteria as described in Section 2.3.1.5. Average loss is calculated as the sigmoid cross entropy, averaged across all evaluated data points (See Equation 2.5). The model was trained for 2000 iterations, without early stopping.

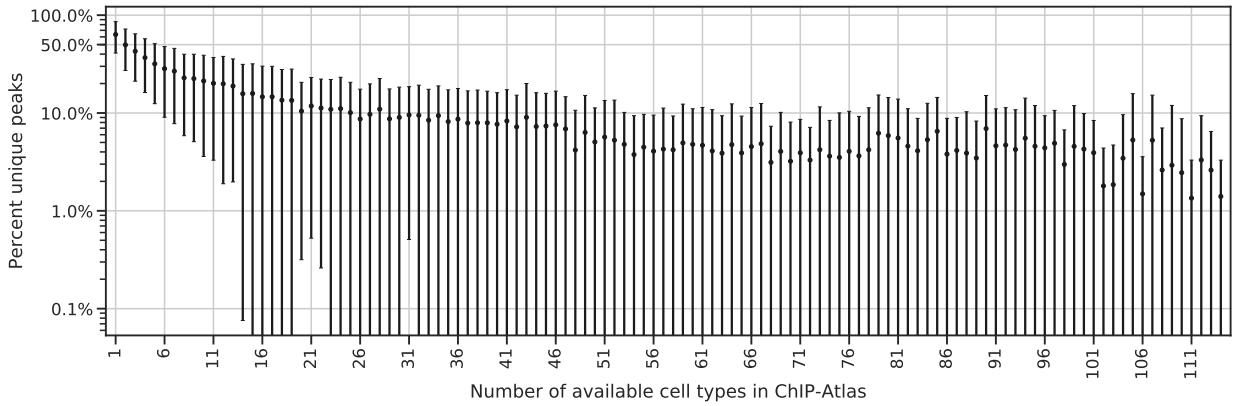


Figure 2.5: Weighted means and standard deviation of percent of unique peaks observed in a cell type as the number of available cell types for a given ChIP-seq target increases. Means and standard deviations are weighted inversely proportional to the number of data points for a given ChIP-seq target. ChIP-seq targets include transcription factors, histone modifications, chromatin accessibility, chromatin modifiers, and histones from called peaks in the ChIP-Atlas database [158].

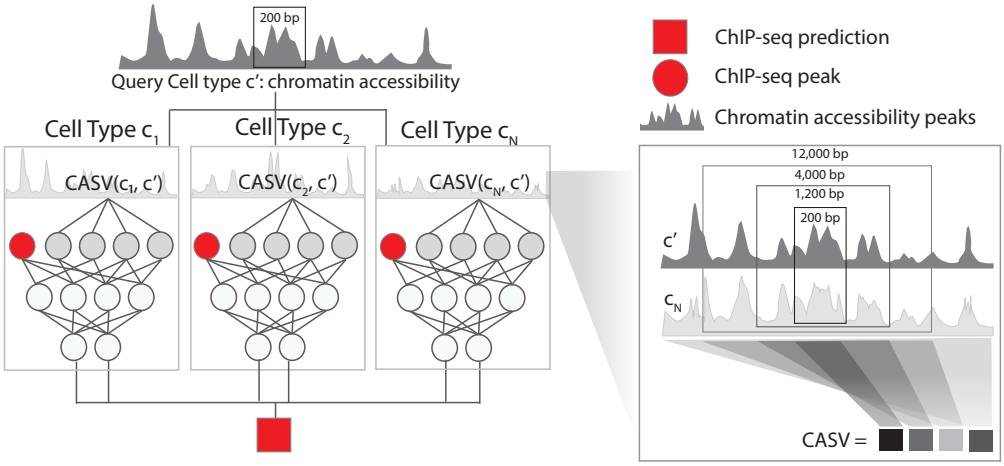
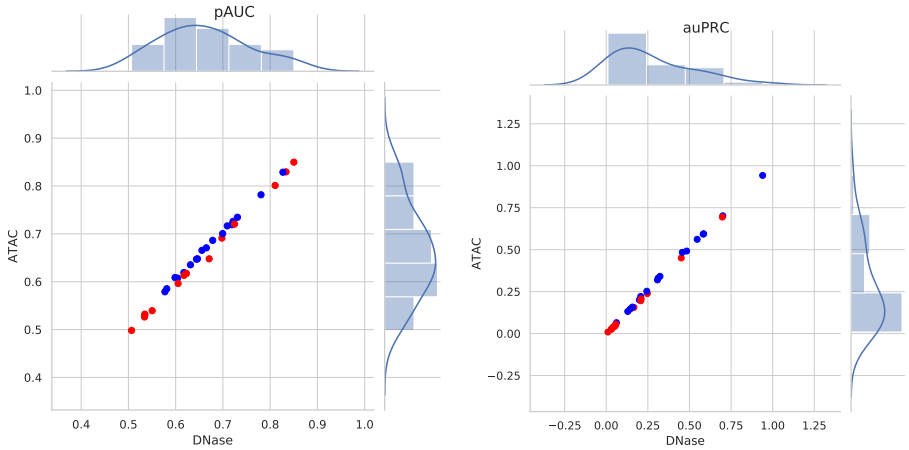
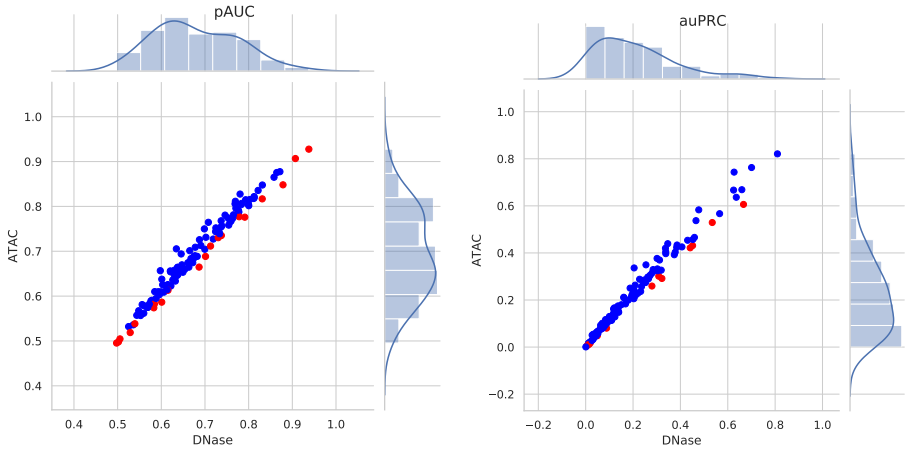


Figure 2.6: Schematic of Epitome for a single ChIP-seq target. Features for each cell type include ChIP-seq peaks at a genomic locus and the chromatin accessibility similarity vector (CASV), which compares the chromatin accessibility of each reference cell type to the query cellular context. The model outputs ChIP-seq peak probabilities for the query cellular context.



(a)



(b)

Figure 2.7: Epitome performs comparably when training models using DNase-seq and evaluating on a new cell line using ATAC-seq. (a) Comparative pAUC (5% FPR) and auPRC performance of 33 TFs when predicting genome wide binding in the A549 cell line using an Epitome model trained using DNase-seq. x axis shows pAUC (left) and auPRC (right) using ENCODE A549 DNase-seq during evaluation, and y axis shows pAUC (left) and auPRC (right) using ENCODE ATAC-seq during evaluation. Blue indicates TFs that perform better when predicted using ATAC-seq data during evaluation. Red indicates TFs that perform better when predicted using DNase-seq data during evaluation. (b) Comparative pAUC (5% FPR) and auPRC performance of 128 TFs when predicting genome wide binding in the K562 cell line using an Epitome model trained using DNase-seq. x axis shows pAUC (left) and auPRC (right) using ENCODE K562 DNase-seq during evaluation, and y axis shows pAUC (left) and auPRC (right) using ENCODE ATAC-seq during evaluation.

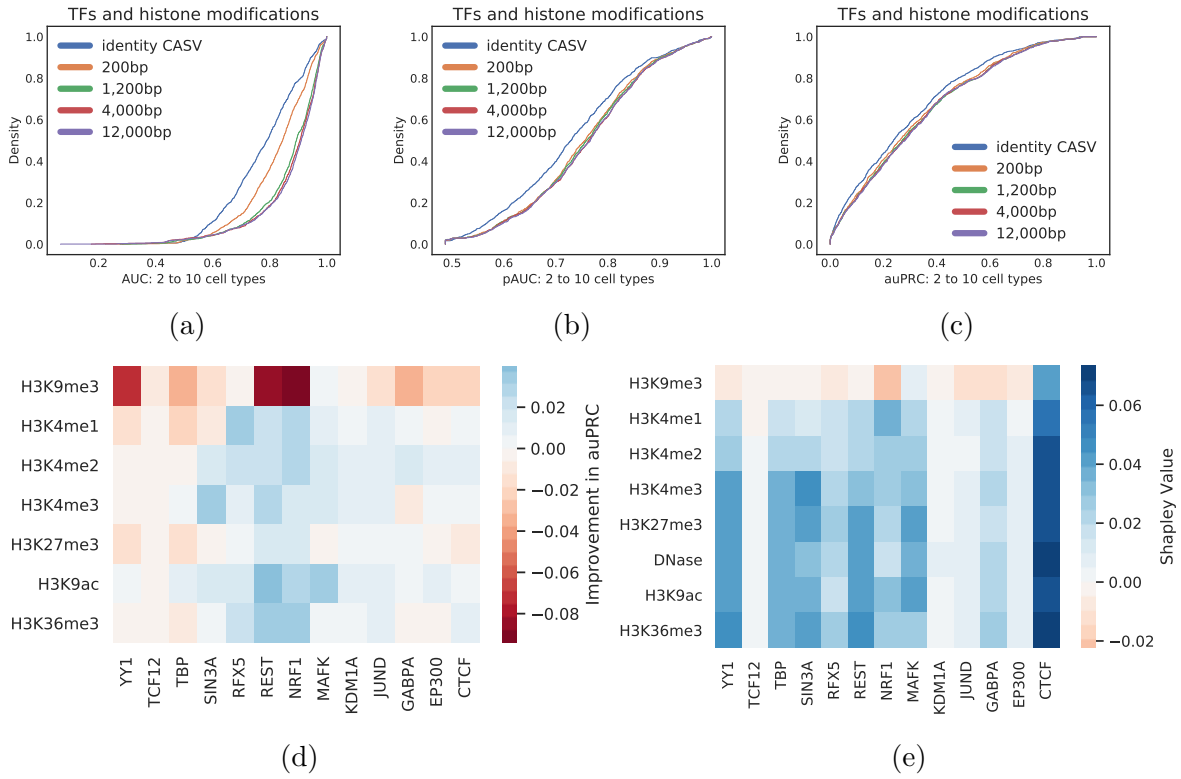


Figure 2.8: Considering wide genomic contexts and multiple epigenetic signals to compute cell type similarity improves model performance. (a) Cumulative distribution functions (CDFs) of Epitome performance in terms of area under the receiver operating characteristic curve (AUC) for TFs and histone modifications in Epitome models trained on 2 to 10 cell types. CDFs of Epitome performance in terms of pAUC (b) and auPRC (c) as various DNase-seq window sizes are considered for computing the chromatin accessibility vector (CASV). Only DNase-seq is used to compute cell type similarity in the CASV. DNase-seq window sizes considered include the identity CASV, 200bp, 1,200bp, 4,000bp, and 12,000bp around a peak of interest. Only models training on less than 10 cell types were considered. Identity CASV implies that no CASV is used in Epitome. (d) Difference in auPRC for 13 TFs when Epitome uses a single histone modification and DNase-seq in the CASV, compared to performance when only DNase-seq is used in the CASV. All 200bp regions on chromosome 7 that have at least one ENCODE epigenetic event were evaluated, where positive include 200bp regions that overlap a ChIP-seq peak for the ChIP-seq target evaluated, and negative regions are all 200bp regions not overlapping a ChIP-seq peak. (e) Shapley values of seven histone modifications and DNase-seq demonstrating their contribution of auPRC performance of 13 TFs when incorporated into the CASV.

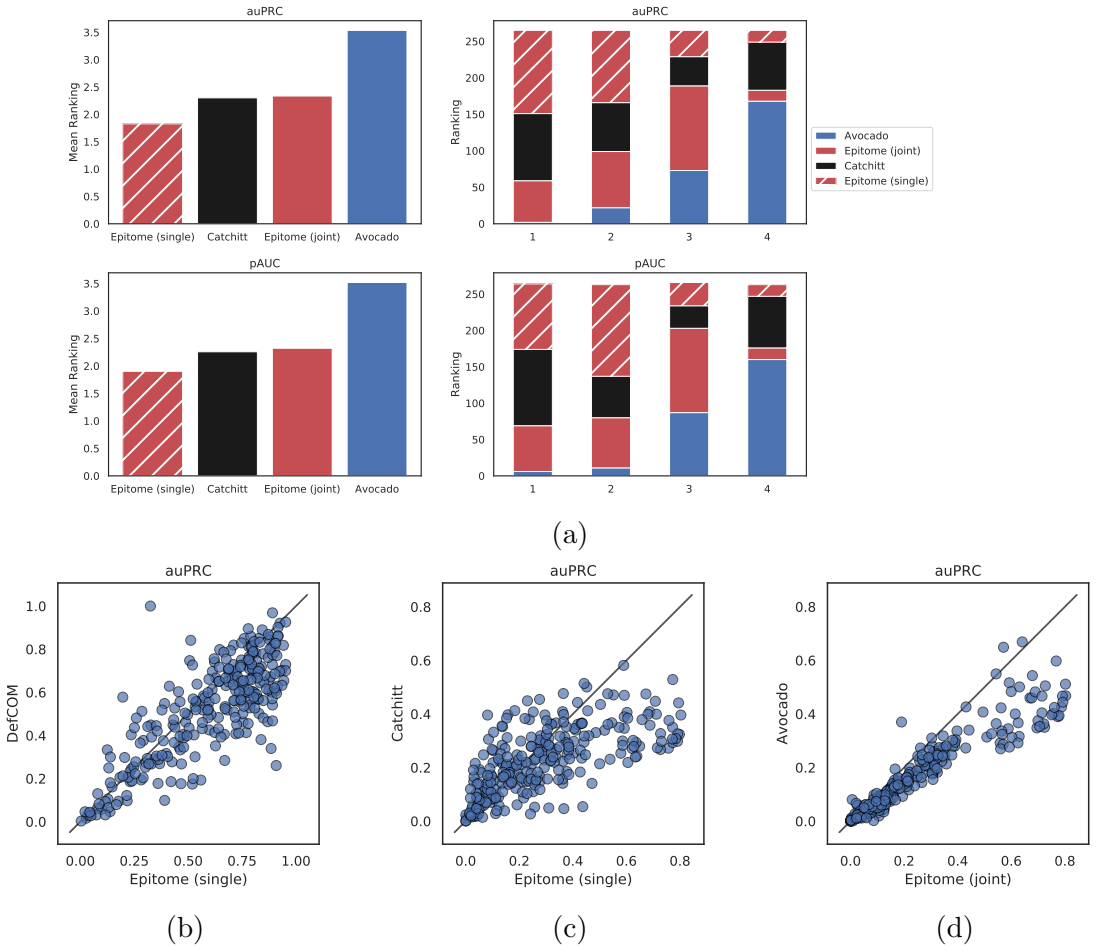
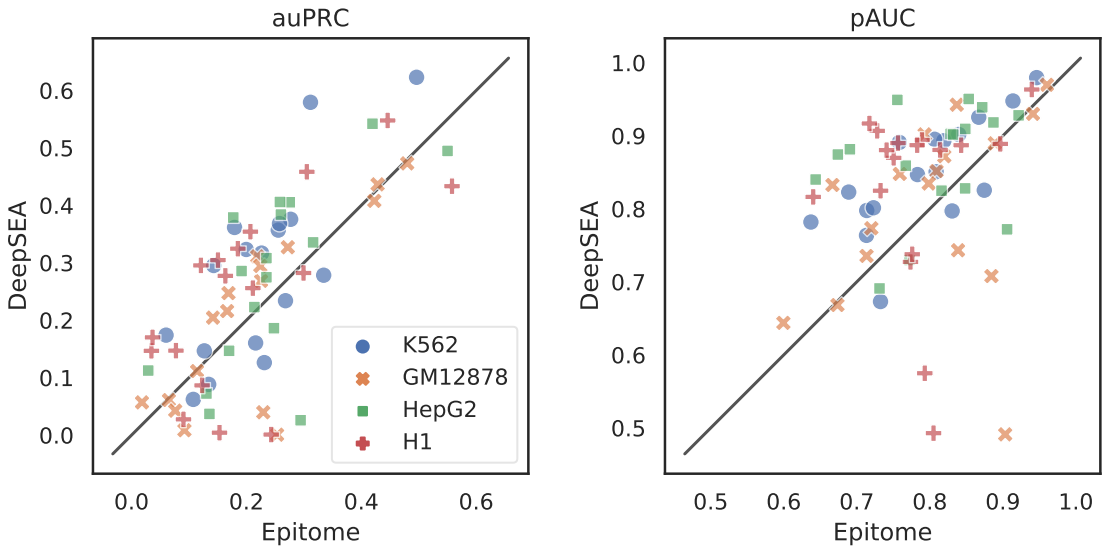


Figure 2.9: Comparison of methods for predicting transcription factor binding sites (TFBS) for 77 TFs and chromatin modifiers in 40 primary cells, cell lines, and tissues from ENCODE. TFBS were predicted on all 200bp regions on chromosomes 8 and 9 that overlap at least one binding site in at least on of the 40 cell types considered. (a) Frequency at which each method obtains a rank for predicting TFBS across 77 transcription factors and chromatin modifiers in 40 held out cell lines, tissues, and primary cells, totaling 264 comparisons. Evaluated methods include Avocado [186], Catchitt [100], a joint Epitome model, and single Epitome models, where each TF is trained separately. (Left) Mean pAUC (5% FPR) and auPRC ranking for each method. (Right) Frequency at which each method obtains a rank based on pAUC and auPRC. (b) Scatter plots comparing auPRC between Epitome and DeFCoM single models. Only regions overlapping motifs specific to the TF being evaluated were considered. (c) Scatter plots comparing auPRC between Epitome and Catchitt single models. (d) Scatter plots comparing auPRC between Epitome and Avocado joint models.



(a) (b)

Figure 2.10: Performance metrics of Epitome and DeepSEA for predicting ChIP-seq peaks for 17 transcription factors on chromosomes 8 and 9 in four held out cell lines, resulting in 68 comparisons. Four held out cell lines include K562, GM12878, HepG2, and H1. Transcription factors compared include: CEBPB, CHD2, CTCF, EP300, GABPA, JUND, MAFK, MAX, MYC, NRF1, RAD21, REST, RFX5, SRF, TAF1, TBP, and USF2. (a) auPRC and (b) pAUC (5% FPR) scores for Epitome and DeepSEA.

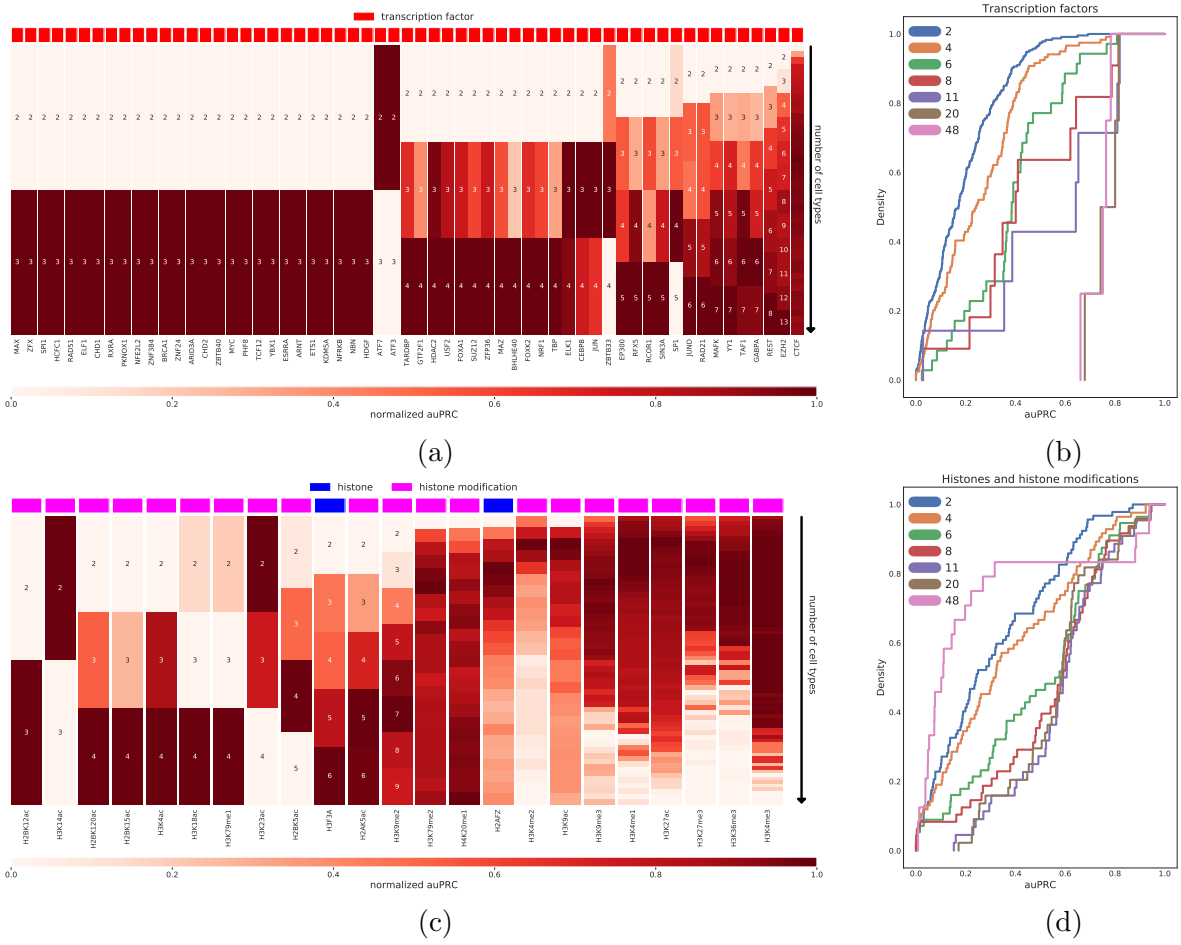
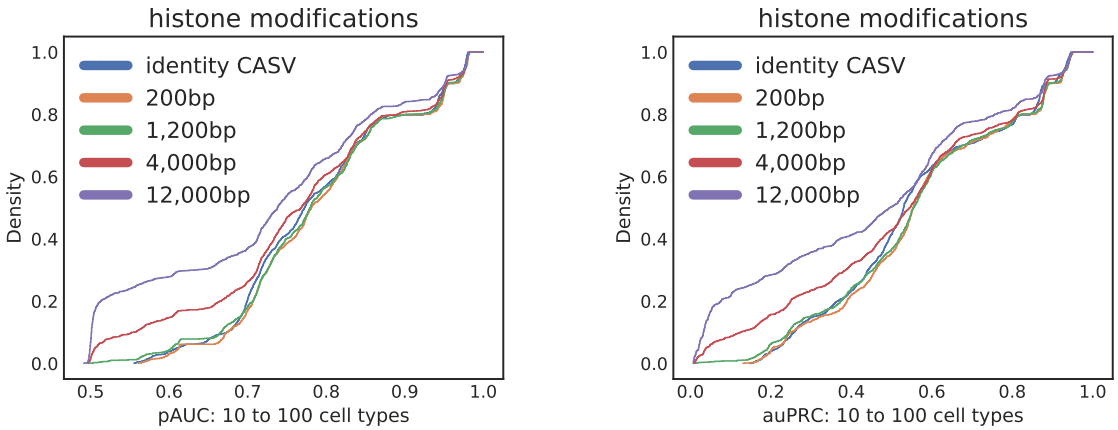


Figure 2.11: The number of cell types selected to train an Epitome model changes predictive performance of transcription factors, histones, and histone modifications. (a) Normalized mean auPRC of 59 transcription factors in heldout chromosome 7 as more cell types are incorporated into Epitome for training. For each set of reference cell types considered for a given TF, mean auPRC was calculated across four models with different combinations of training and validation cell types. The number of training cell types considered ranges from 2 to 48 cell types. This range is dependent on the availability of reference cell types for a given transcription factor in ENCODE. (b) Cumulative distribution function (CDF) of auPRC performance of 59 transcription factors in heldout chromosome 7. (c) Normalized mean auPRC of 23 histone modifications and histones in heldout chromosome 7 as more cell types are incorporated into Epitome for training. Mean auPRC was calculated across four models with different combinations of training and validation cell types. The number of training cell types considered ranges from 2 to 84 cell types. This range is dependent on the availability of experiments for a given histone modification or histone in ENCODE. (d) Cumulative distribution function (CDF) of auPRC of 23 histone modifications and histones in heldout chromosome 7.



(a) (b)

Figure 2.12: Considering genomic contexts of various sizes to compute cell type similarity in the CASV affects performance of transcription factors (TFs) and histone modifications. Various DNase-seq window sizes are considered for computing the chromatin accessibility vector (CASV). Only DNase-seq is used to compute cell type similarity in the CASV. DNase-seq window sizes considered include no DNase-seq, 200bp, 1,200bp, 4,000bp, and 12,000bp around a peak of interest. (a),(b) CDFs of Epitome performance for histone modifications in Epitome models trained on more than 10 cell types. Performance was measured in (a) partial area under the receiver operating characteristic curve (pAUC) (5% FPR) and (b) area under the precision recall curve (auPRC).

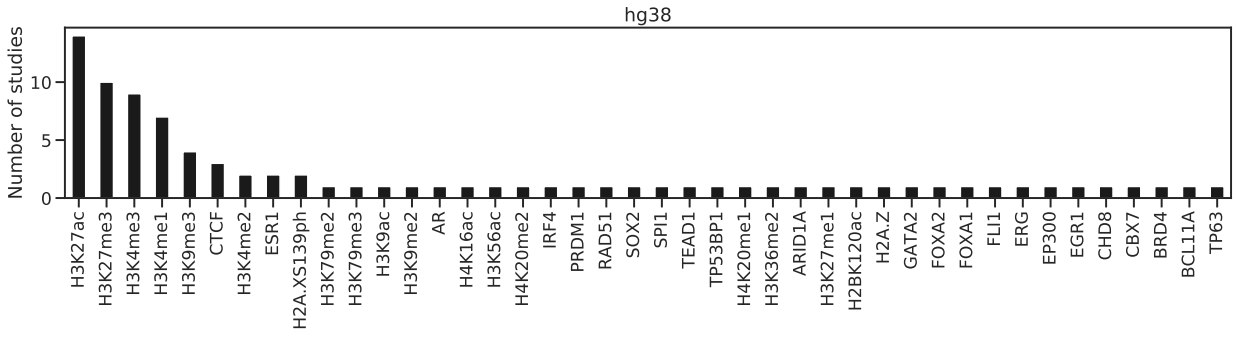


Figure 2.13: Frequency at which a given ChIP-seq target is present in a study included in ChIP-Atlas [158] that contains at least two ChIP-seq experiments. Only includes studies aligned and processed under the hg38 genome.

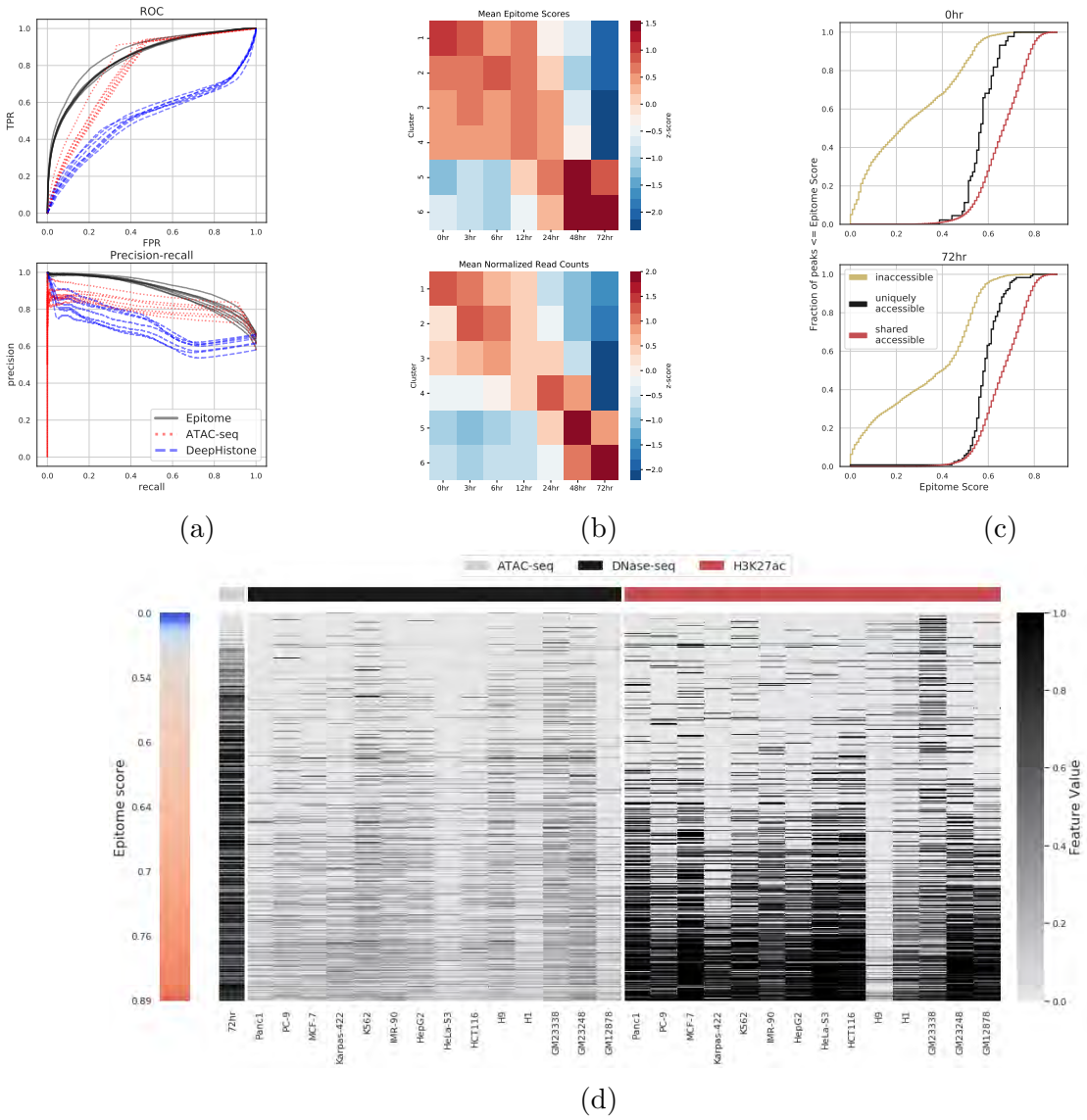
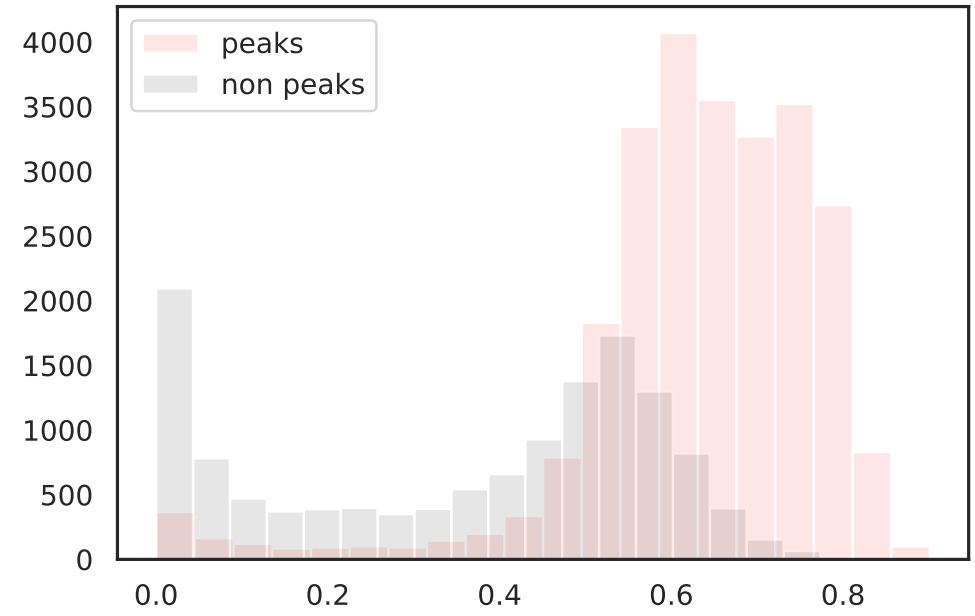
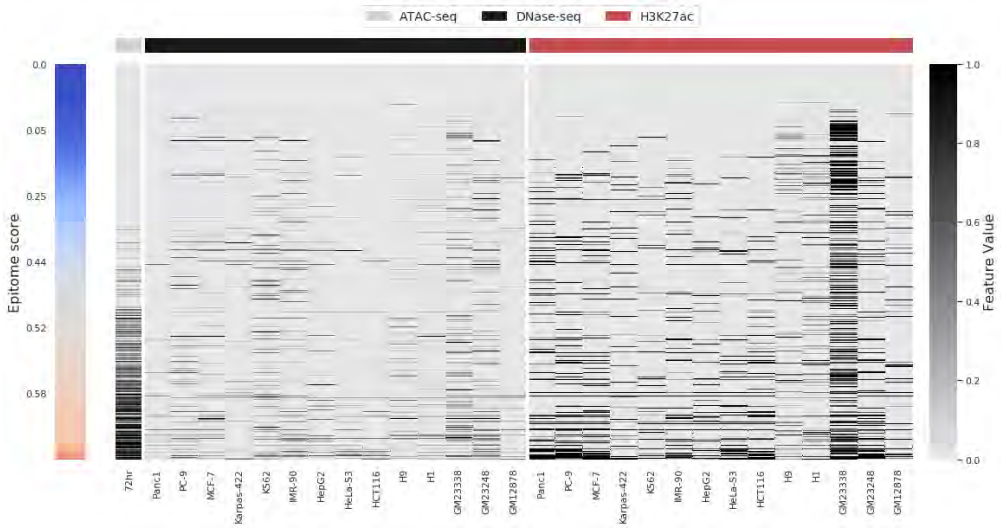


Figure 2.14: Epitome detects differential H3K27ac across seven time points in neural differentiation. (a) ROC and PR curves for predictions of H3K27ac peaks using three methods at seven time points of neural differentiation. (b) (Top) Mean Epitome scores of H3K27ac peaks across seven time points in six clusters from 2,400 temporal peaks [83]. (Bottom) Mean normalized H3K27ac read counts across seven time points in six clusters. Rows are standardized. (c) CDFs of Epitome scores for H3K27ac peaks at 0hr and 72hr in regions that are uniquely accessible to a timepoint (black), are inaccessible for a timepoint (yellow) and have shared accessibility across all timepoints (red). respectively. (d) Heatmap of features used by the Epitome model for 25,762 genomic regions containing H3K27ac peaks at 72hr. ATAC-seq column, labeled in grey, indicates presence of absence of ATAC-seq peaks in the 72hr time point. Color bar on left represents Epitome scores, where blue represents instances of false negatives and red represents instances of true positives.



(a)



(b)

Figure 2.15: (a) Epitome predictions of H3K27ac peaks at 72hr after neural induction for peak and nonpeak regions. (b) Heatmap of features used by Epitome for 13,248 regions that do not contain H3K27ac peaks at 72hr. Color bar on left represents Epitome scores, where blue represents true negatives and red represents false positives.

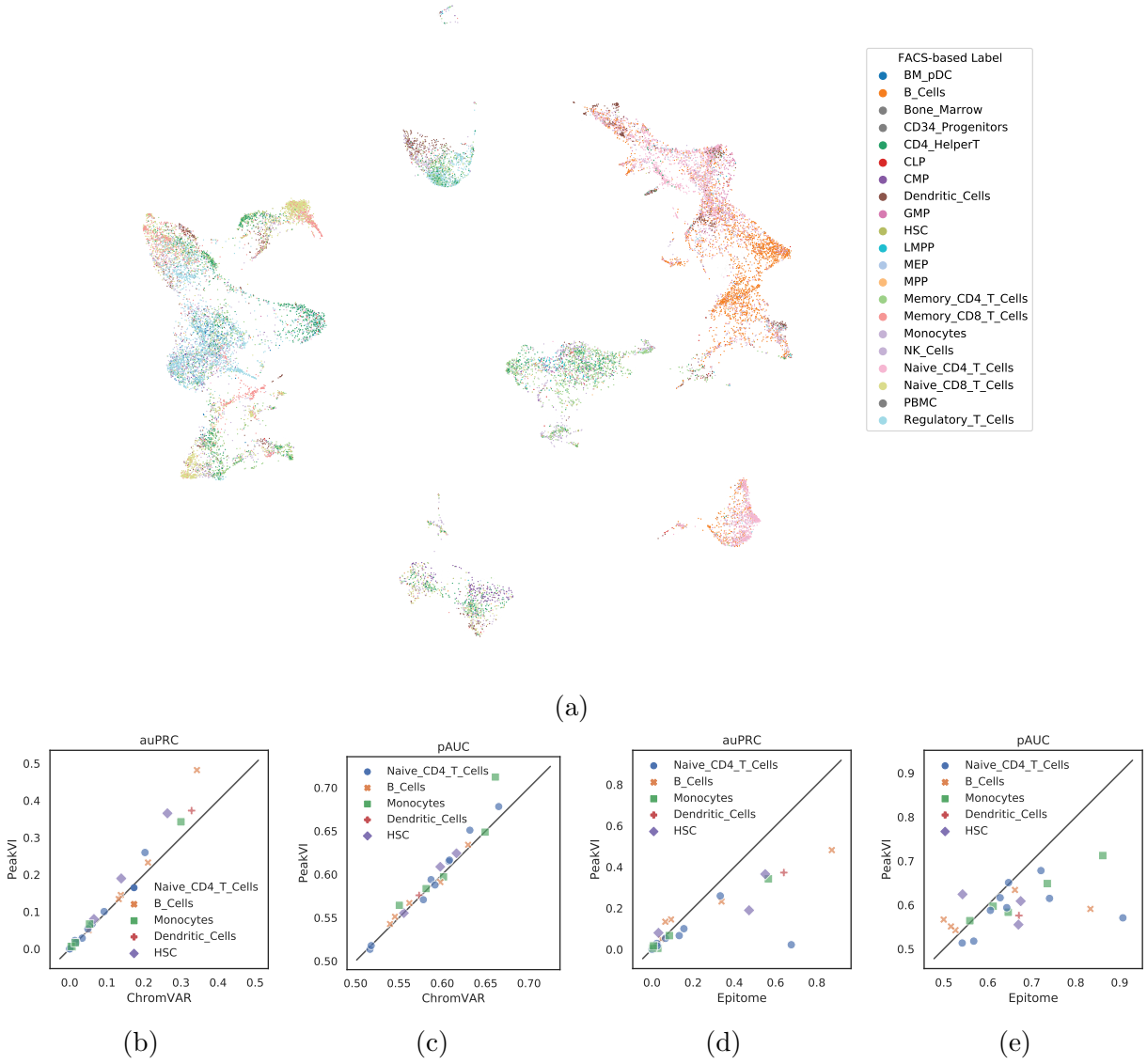


Figure 2.16: (b) Area under the precision recall (auPRC) and (c)partial area under the ROC (5% FPR, pAUC) for prediction of 17 TFs from ChIP-Atlas on pseudo-bulk populations identified from scATAC-seq [179]. Comparison of two methods: motif overlaps using PeakVI posteriors and chromVAR background corrected scATAC-seq fragments. (d) auPRC and (e) pAUC for prediction of 17 transcription factors (TFs) from ChIP-Atlas on pseudo-bulk populations identified from scATAC-seq [179]. Comparison of two methods: motif overlaps using PeakVI posteriors and TFBS predictions from scEpitome.

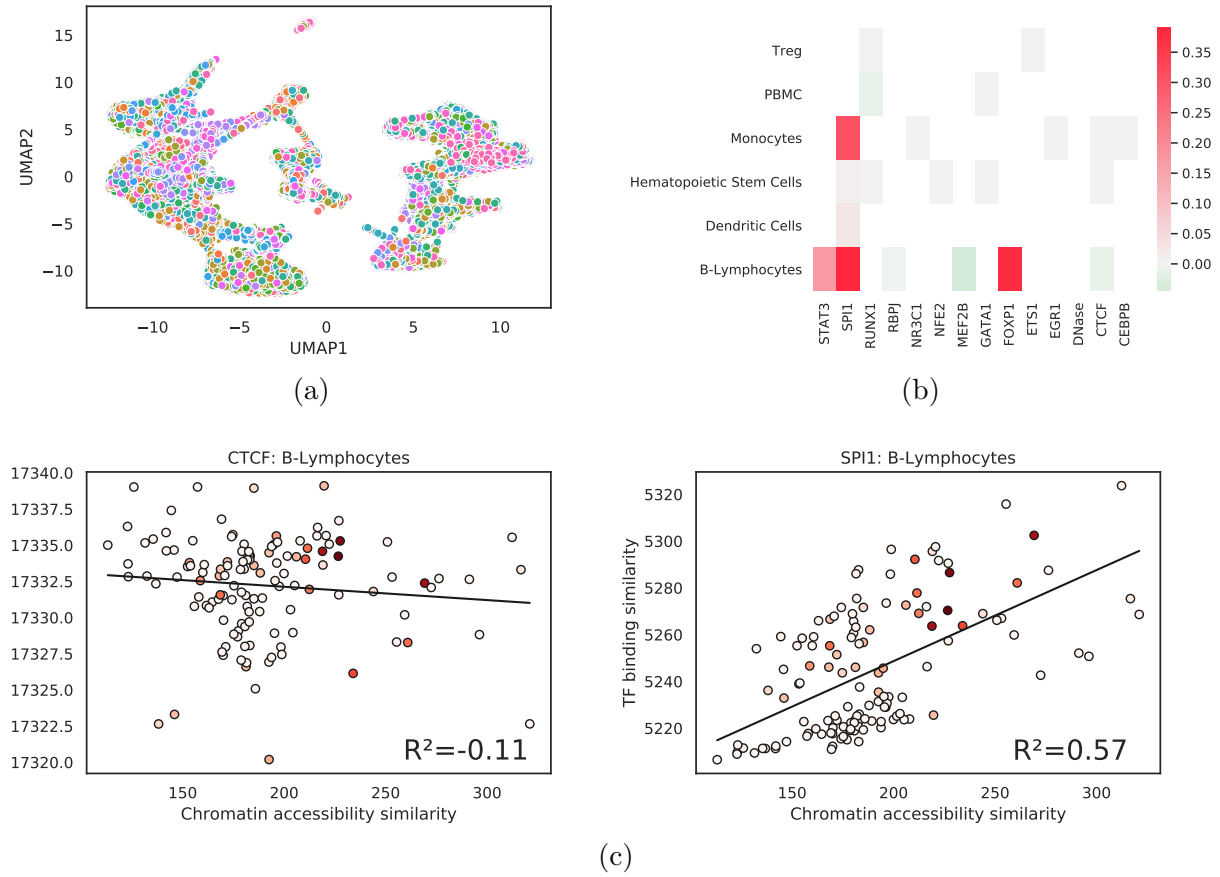


Figure 2.17: scEpitome predicts transcription factor binding sites in microclusters computed from PBMC and bone marrow scATAC-seq [179]. (a) Projection of UMAP, computed from PeakVI latent space, colored by 128 microclusters computed with VISION [40]. (b) Slope of the curve for chromatin accessibility similarity (comparing bulk DNase-seq to mean scATAC-seq across microclusters) vs similarity in TF binding. TF binding similarity is calculated as the dot product between binary bulk ChIP-seq peaks and scEpitome predictions. Slope is calculated from linear least-squares regression. (c) Example scatter plots of chromatin accessibility similarity and TF binding similarity between microclusters and bulk datasets. CTCF and SPI1 from B-lymphocytes. Similarity is calculated as dot product between bulk binary peaks and PEAKVI or scEpitome probabilities. Microclusters are colored by number of cells in each cluster that are labeled by MACS as B-lymphocytes. R represents correlation coefficient.

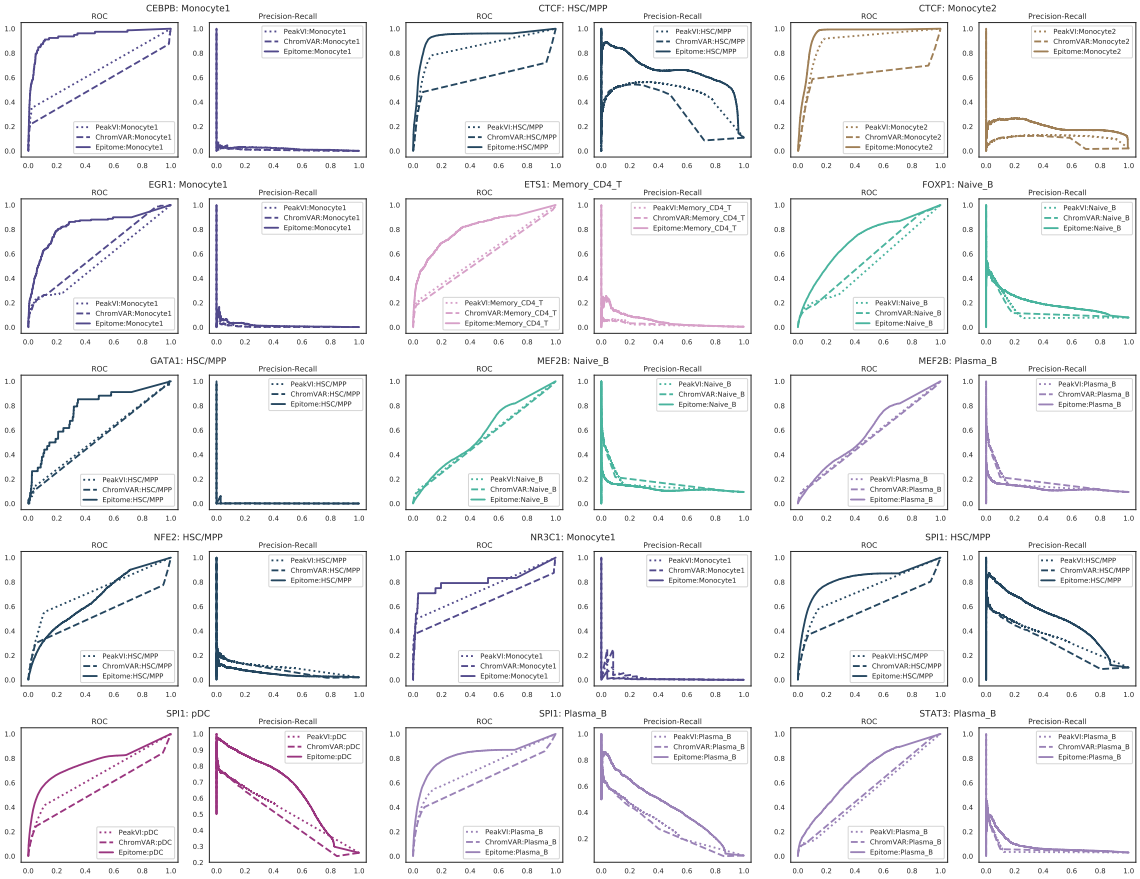


Figure 2.18: Example ROC and precision-recall plots for predicting TFBS from scATAC-seq in peripheral blood mononuclear cells (PBMCs). Methods compared include Epitome, a motif overlap analysis using bias corrected fragments, as used by ChromVAR [181], and motif overlap using PeakVI posteriors. Bulk ChIP-seq was used as ground truth in 5 PBMC cell types, including, monocytes, dendritic cells, HSCs, B cells, and naïve CD4(+) T cells.

Chapter 3

Computational tools for analysis of genomic sequencing data

If you have ever worked on a machine learning or data science project, you may have heard of the saying "garbage in, garbage out". This saying means that the quality of your analysis, or the quality of your output from a predictive model, is limited by the quality of the data you put in to it. It has been observed that the accuracy of machine learning models is correlated with the cleanliness and clarity of data used as input [206]. Therefore, it is commonly known in the data science community that when learning from a data set of interest, data cleaning and processing are the most time consuming components of the pipeline [223]. Indeed, an Epitome model, discussed in Section 2.3.1, used for predicting TFBS can train in 14 to 50 minutes. However, downloading, normalization, and consolidation of training data from ENCODE takes up to 23 hours!

Because of this fact, much of the time and energy that went into this work involved cleaning and processing various genomic data sets. In this chapter, we discuss three tools, or extensions of tools, that we have built to enhance data processing and analysis. We have specifically focused on making these tools reusable and reproducible so they can be utilized by other members in the lab, or by the community in general.

3.1 Computational pipelines for quantification of endogenous retroviruses

Endogenous Retroviruses (ERV), otherwise known as long terminal repeat (LTR) retrotransposons, are a group of retroviruses that make up 8% of the human genome [36]. Although many ERVs are defective for infection, these repetitive regions contain regulatory regions, such as TFBS, that are required for proviral transcription [1] [211]. Because of this, several

[1] proviruses are viral genomes that are integrated into the DNA of a host cell

pathways have evolved to suppress ERVs, such as DNA methylation and the establishment of H3K9me3 [66].

In a recent study, Groh et.al. showed that a TF called MORC3 binds ERV sequences in mouse embryonic stem (ES) cells, silencing ERVs [67]. As discussed in Chapter 5, we also investigate the repressive characteristics of MORC3, albeit in human monocytes. Thus, we wanted to check whether MORC3 was repressing ERVs in human monocytes, as it was in mouse ES cells. To determine this, we collected RNA-seq data from wild type (WT) and MORC3 knock out (KO) monocytes, as described in Chapter 5. We then could compare the expression levels of ERVs between both conditions to determine whether ERVs had significantly increased expression in the KO.

In order to build a pipeline to quantify ERVs from RNA-seq data, we extended a project called SeqTools, which was originally introduced in the work of James Kaminski [97]. SeqTools is an object-oriented system designed for end-to-end processing of high-throughput data, including ATAC-seq, ChIP-seq, and RNA-seq. Through SeqTools, and the help of Hector Roux de Bézieux in Sandrine Dudoit's lab at UC Berkeley, we implemented a pipeline to quantify ERVs from RNA-seq samples. This pipeline requires three steps:

1. Alignment of reads to the transcriptome
2. Quantification of family-level expression levels for ERVs
3. Detection of differential expression of ERV families

We used the STAR aligner to align reads to the transcriptome [42]. Repeats containing the location of ERVs were downloaded from the UCSC table browser [98]. Because ERVs are repetitive, it is difficult to determine the exact origin of sequences that align to ERVs. Because of this, we quantify expression levels for ERV families, instead of individual ERVs. We used featureCounts to count reads in the transcriptome that overlapped ERV families [120]. Finally, counts of reads overlapping ERV families were compared across samples to determine differential expression of ERVs between conditions using DeSeq2 [126].

3.2 Computational pipelines for processing CUT&RUN

As discussed in Section 1.1.1, ChIP-seq is a commonly used method to profile TFBS and histone modifications. In this method, cells are first treated with formaldehyde for protein-DNA crosslinking. Chromatin is then fragmented, and antibodies specific to the protein of interest is added. Finally, antibody bound sequences are extracted and sequenced [196]. However, numerous problems with ChIP-seq have driven the search for more efficient profiling techniques. These issues include low extraction efficiency of protein-DNA complexes, and "hyper-ChIPable" regions of the genome that have artificially high levels of enrichment [208].

With these problems in mind, Skene and Henikoff developed Cleavage Under Targets and Release Using Nuclease (CUT&RUN) that improve over ChIP in several ways. First of all, DNA in starting cells do not need to be fragmented prior to extraction, and thus mimic natural protein-DNA interactions better than that of ChIP. Second, CUT&RUN requires only a fraction of the sequencing depth of ChIP, allowing for smaller starting cell numbers. Finally, CUT&RUN as been shown to more accurately identify TFBS than ChIP-seq [193].

Due to these benefits, we and collaborators collected CUT&RUN data for various histone modifications and the TF MORC3, discussed in Section 5. Here, we briefly describe our extension of SeqTools to process CUT&RUN.

To process CUT&RUN data, we modify the ChIP-seq pipeline implemented in SeqTools as a basis for processing, with two crucial changes. The original ChIP-seq processing pipeline includes three stages, including (1) alignment of reads to a reference, (2) identification of peaks, and (3) combining peaks across all samples to define a peak universe. The first regards alignment. Similar to the ChIP-seq pipeline, we use bowtie2 [112] to align reads to the reference genome. However, we enable dovetail alignment for CUT&RUN to accept paired-end reads when there is overlap between mate pairs, which is infrequently encountered in CUT&RUN experiments [241]. The second change made to the pipeline is the choice of peak caller used to identify peaks in a given sample. While MACS2 [239] is used to identify peaks in ChIP-seq samples, this method uses a Poisson model determine enrichment of regions over a background. However, because CUT&RUN has lower background signal, such methods can have reduced precision when used on CUT&RUN [135]. Due to this issue, we utilize SEACR in SeqTools, a peak caller specifically designed for CUT&RUN data that is model free and data-driven [135].

3.3 Visualization of large-scale genomic sequencing datasets

As computational biologists, visualization is one of the primary ways which we are able to cope with the overwhelming data complexity we deal with on a day-to-day basis [162]. The transition from Sanger to high-throughput DNA sequencing has led to a dramatic expansion of genomic sequencing datasets [191], driving diverse and integrative studies such as TCGA, TOPMed, and ENCODE, which contain more than 300 TB of sequencing data [209, 133, 35, 1, 131]. Consequently, these massive datasets are being mined to develop and validate hypotheses for understanding changes in activity of regulatory regions across cellular contexts, target treatment interventions, and drive discoveries in precision health [184]. One such example of this is provided in Chapter 2, where we introduced two prediction methods that leverage data from the ENCODE consortium to learn about changes in TF binding and histone modifications across cellular contexts [143, 144]. However, this pileup of data does not stop at public consortiums. As demonstrated in Chapters 4 and 5, personal studies are collecting multi-omics sequencing datasets with ever-more replicates and biological

conditions. In Morrow et. al. [145], we introduce a tool called Mango, which consists of a Jupyter notebook and genome browser component that supports visualization of large genomic datasets in a genome browser-like format. Mango removes scalability and interactivity constraints of existing interactive tools by leveraging multi-node compute clusters to allow interactive analysis over terabytes of sequencing data. This section discusses Mango and its extensions, and is modified from Morrow et al. [145].

Existing tools for evaluating genomic sequencing datasets either support abstractions for processing multi-terabyte datasets or interactive analysis, but do not provide both. Tools like bioconductor [80] support interactive analysis on sequencing data by providing a flexible programmatic environment for quality assessment, ad hoc visualization, event detection and summarization of data. Other tools, such as IGV, Savant, and IOBIO support static visualizations of sequencing samples at single sample resolution [174, 53, 137, 194, 56]. The scaling restrictions of existing genomics tools has led to the development of new tools for preprocessing genomic datasets by leveraging distributed platforms such as Apache Hadoop and Apache Spark [180, 157, 156], allowing users to elastically scale compute resources as demanded by dataset size. Although these tools scale to terabyte-sized datasets, they are primarily designed for preprocessing multiple samples using specific algorithms for alignment and variant calling, and do not provide primitives for interactively evaluating and iterating on post-processed results. In industry, proprietary platforms have been developed for interactively visualizing large-scale genomic datasets [171]. As large-scale sequencing provides a lens into the complex genomic biology that drives disease, tools that make it possible to interact with large cohorts are necessary to enable scientists to understand complex patterns in their data.

Mango is an open source genomics visualization platform consisting of a browser and notebook form factor. Both the Mango browser and notebook support visualization of remotely staged genomic datasets consisting of alignments, variants, and features, and are built on Apache Spark [236] and ADAM [156] to scale to terabyte sized genomic datasets. Mango supports ad-hoc exploration of terabyte sized datasets on private clusters and in cloud storage, removing the requirements of costly dataset transfers from centralized repositories like dbGaP [238]. Mango supports human-in-the-loop analysis found in single-node tools by providing a flexible programmable interface in the form of the notebook, and a traditional genome track browser for exploring individual samples. The Mango notebook allows users to programmatically manipulate and visualize datasets in a Jupyter notebook environment by providing access to genome track Jupyter widgets and summary visualizations. Meanwhile, the Mango browser provides traditional track views for variants, features, and variants, while allowing users to remotely access datasets. Architecture for the Mango notebook and browser are shown in Figure 3.3.

The Mango notebook supports querying, exploring, and summarizing alignment and variant datasets in a Jupyter notebook environment [106], matching the interactivity of genomics analytics tools such as Bioconductor while removing the static constraints of genome browsers. The Mango notebook also supports traditional scrollable genome track visualizations through Jupyter widgets built on pileup.js [218] for interactive visualization of pileups,

variants and features. These widgets can be run on local standard genome file formats such as bigBed, BAM and VCF file formats, but can also be combined with Apache Spark to visualize remotely-staged files in parquet formats. Figure 3.2a demonstrates sample visualizations for analysis of genome coverage, run on 10 TB of 100 genomes from the Simons Genome Diversity dataset [131]. Visualization results from functions provided through the Mango notebook are returned as Matplotlib objects [81], allowing users to modify and update plots without recomputing on the entire dataset. Programmatic querying and filtering of datasets can be performed through Apache Spark SQL queries and manipulations on Apache Spark resilient distributed datasets (RDDs). Figure 3.2b shows how pythonic filters can be combined with visualization of data from the 1000 Genomes Project [1]. Finally, we support visualization of deeply sequences samples. Figure 3.2c shows track visualization of 780 GB of alignment data from a seven sample subset from the Simons Genome Diversity dataset using the widgets.

While the Mango notebook provides a programmable interface for exploring genomic sequencing datasets, the Mango browser provides a graphical user interface (GUI) for visualizing subsets of cohorts. The Mango browser supports analysis and exploration of genomic reads, variants, and features at user-defined loci, supporting functionality for quality analysis and local correspondence between samples. The Mango browser supports functionality implemented in commonly used browsers such as IGV, IGB, JBrowse and Savant [174, 56, 53, 194], while additionally allowing users to access remotely staged genomic sequencing samples from dataset repositories. The Mango browser stores data in efficient interval-based structures, or Interval RDDs [142] to allow fast and consistent response times of overlapping genomic range queries on cached datasets, compared to Apache Spark's default RDD. Figure 3.2d shows example visualizations of the home screen and browser view of a remotely staged alignment file in the Mango browser.

3.4 From large to small: visualization in a flexible Jupyter environment

Since the publication of Morrow et.al. [145], we have found through our own epigenetic workflows that dataset size is not always the main constraint when it comes to visualization. Often, we process large datasets through a "black box" pipeline, where we are not concerned with visualizing the intermediary results. However, we are more interested in visualizing datasets after processing, whose size is small enough to fit on a single machine. Therefore, tools built on top of Apache Spark [145] result in unnecessary memory and management overhead when we are only interested in visualizing a small, post-processed dataset. Because of this, we have introduced the Mango widgets, a set of stand-alone embeddable widgets that allow visualization of features, variants, and reads in a Jupyter notebook [106]. The Mango widgets are built on Jupyter widgets [95] and pileup.js [218] to provide an embeddable JavaScript-based genome browser.

A key contribution of the Mango widgets is the ability to seamlessly visualize genomic data in the same Jupyter notebook that you are currently analyzing your data in, without requiring loading results into a separate visualization tool. To achieve this, we implemented track configuration options which allow users to customize tracks by color and programmatically zoom in and out of tracks. We additionally implemented an svg saving option, allowing users to programmatically save genomic viewer screenshots within a notebook. These features allow users to integrate genome browsers into existing python pipelines, supporting the ability to programmatically save images of key genomic regions in a single workflow.

Although existing visualization tools such as igv.js [82] support genome visualization in a Jupyter notebook, lack of intermediate data handling limits users to only visualize genomic data stored in genomic file formats that are located within the Jupyter kernel. However, it is often the case that users would like to view genomic data stored in other formats. One example of this is supporting the ability to visualize genomic data stored in dataframes, or numeric feature data stored in numpy matrices [72]. The Mango widgets implement an intermediate data handling layer that allows users to load in both genomic file formats and dataframes. The Mango widgets are able to visualize genomic data stored in both pandas [221] and Modin [140] dataframes, allowing users to seamlessly filter and manipulate data before visualization.

Because genomic datasets can still exceed the limits of pandas dataframes, we provide the option of working with dataframes using Modin [140], a scalable dataframe system that exposes the pandas Application Programming Interface (API) and allows faster manipulation of dataframes in memory. Modin has support for reading in multiple different file types, including compressed and uncompressed files. Internally, Modin has mechanisms for using multiple CPU cores to read in a file, even if the file itself is not split or distributed. In contrast, the standard pandas implementation does not enable any parallelism, as it is only able to read using one processor at a time. Here, we implement parallel readers for genomic file formats to utilize Modin's parallel processing to load in genomic data in a Modin dataframe. Content stored in these dataframes can then be quickly visualized using the Mango widgets.

Example visualizations of genomic features and genotypes supported through the Mango widgets are shown in Figure 3.3. The Mango widgets can be used in both Jupyter notebook and JupyterLab [106].

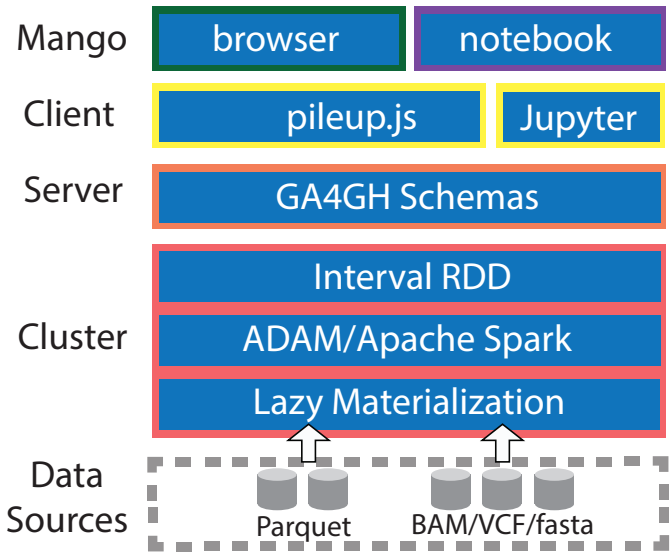


Figure 3.1: Mango architecture, divided into client, server, and cluster components. The Mango browser and notebook are built on ADAM and Apache Spark for fast in-memory data processing. The Mango browser utilizes lazy materialization and Interval RDDs to index and access genomic regions in subsecond latencies [145, 142]. Lazy materialization supports efficient caching on large files, while Interval RDDs support low latency overlapping region queries on genomic data. Both the Mango browser and notebook utilize GA4GH schemas to transfer genomic data serialized in JSON format from the server to the client. The Mango notebook components are accessible through a Jupyter notebook environment [106]. Genome track visualizations for genomic data are rendered using pileup.js [218].

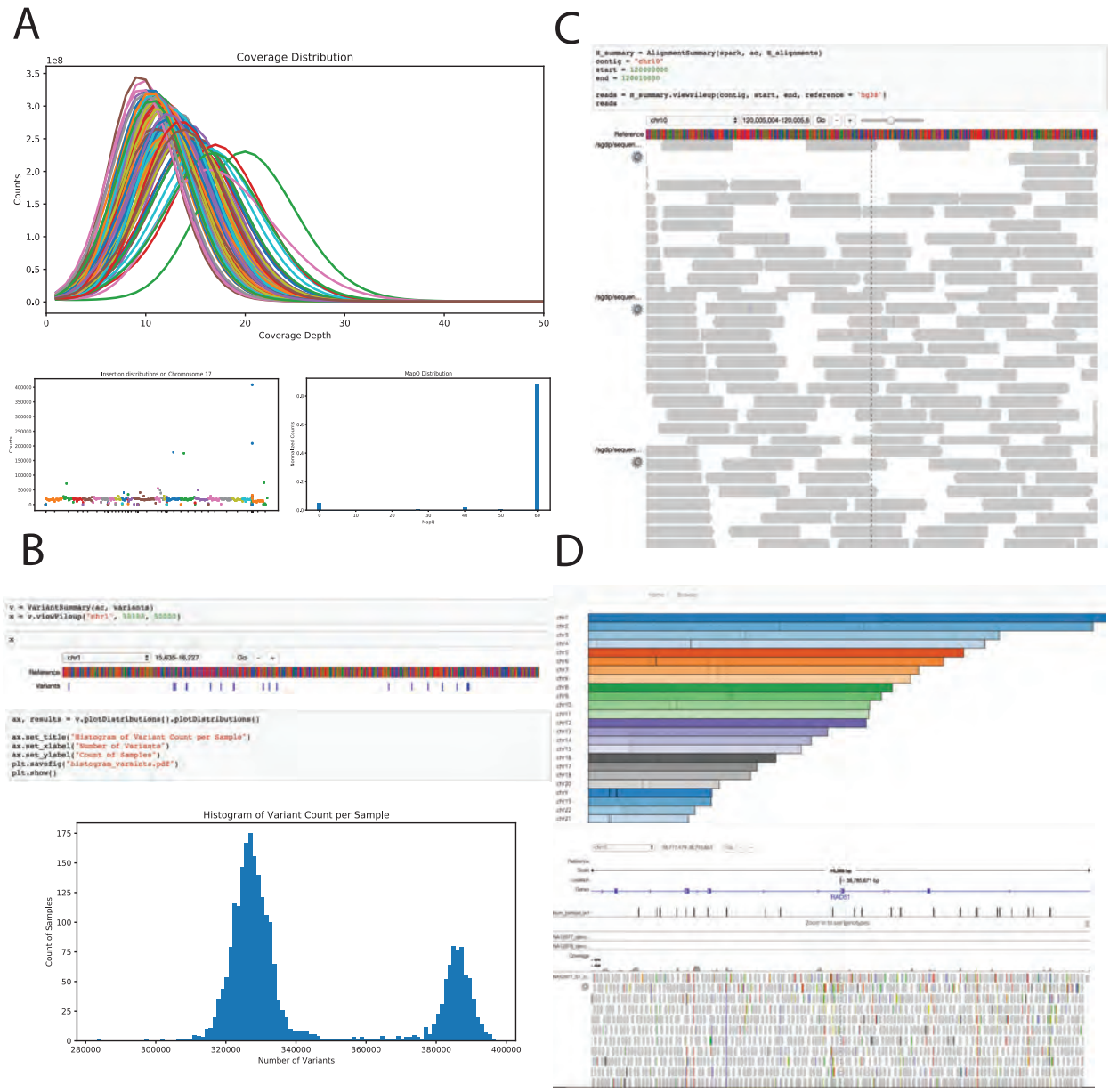
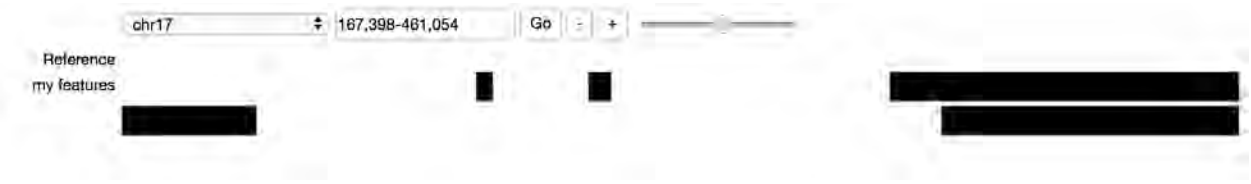
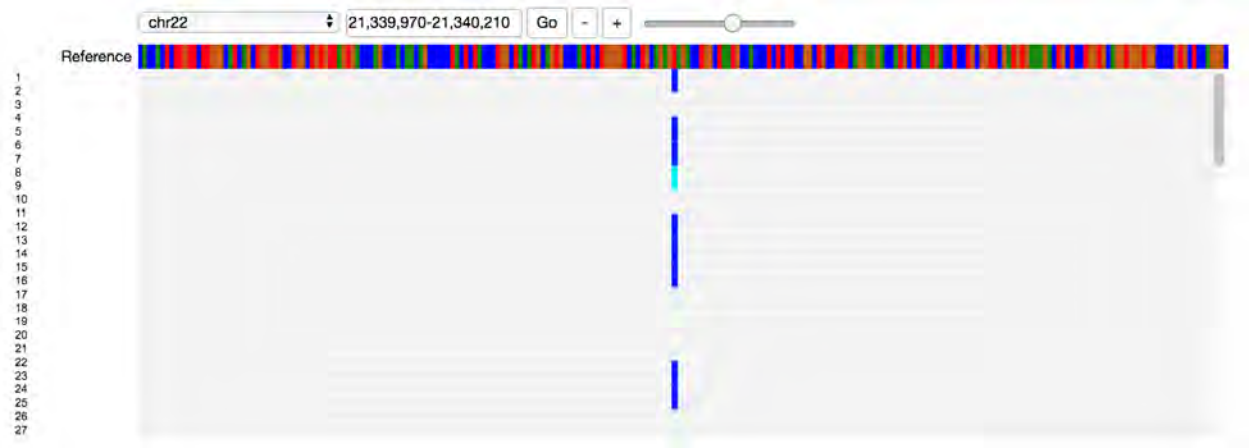


Figure 3.2: Example visualizations from the Mango notebook and the Mango browser. a) Example distributions of genomic sequencing samples in Mango notebook. Coverage distribution of 100 high coverage samples from the Simons Genome Diversity Project. Insertion and mapping quality distributions of a single outlier sample. b) Variant visualizations from chromosome 1 of the 1000 Genomes dataset, including track visualizations of variants in a 1000 bp segment on chromosome 1 and distribution of variants per sample. c) Visualization of the Mango pileup widgets: 780 GB of a seven sample subset from the Simon's Genome Diversity dataset, queried in a Jupyter notebook. d) Home screen and track visualization of sequencing data in the Mango browser.



(a)



(b)

Figure 3.3: The Mango widgets: example screenshots of (a) features and (b) genotypes loaded in a Jupyter notebook.

Part II

Applications

Chapter 4

Transcriptomic and epigenetic memory retention of reprogrammed CD8(+) T cells

In these next two chapters, we explore two applications in which the background, methodology, and pipelines discussed in Chapters 1-3 are applied to characterize the epigenome in multiple cellular contexts. In this chapter, we gather measurements of the transcriptome and chromatin accessibility to characterize CD8(+) T cells, which provide the stone of our analysis. In particular, we explore the extent to which regulatory elements in CD8(+) T cells are maintained after reprogramming to induced pluripotent stem cells (iPSCs), and its application in adoptive T cell therapy. This work is a joint collaboration with the Blazar Laboratory at the University of Minnesota. While experimental work was performed by the Blazar Laboratory, this chapter focuses on the computational analysis we performed for this project.

4.1 Background

T cells play a central role in acquired immunity with the capability of recognizing and responding to foreign antigens, pathogens and damaged cells [89]. Understanding and harnessing key properties of T cell biology has led to the revolutionary clinical application of T cells to treat life-threatening conditions such as cancer [43, 49, 76, 176, 23]. The forced expression of anti-tumor reactive T cell receptors and chimeric antigen receptors (CAR) for the adoptive T cell transfer has improved response rates, especially for those with hematological malignancies [22, 151, 154, 161, 177]. However, more work is needed to improve the durability of these evolving T cellular therapies.

A robust early T cell response to a foreign antigen or pathogen results in a burst of short-lived T effector cells (T_{EFF}) that is followed by a contraction phase. The remaining T cells enter a short- or long-term memory phase. Long-term immunologic memory is essential

for productive immune surveillance. In both mice and humans, CD4(+) and CD8(+) T cells can become long-lasting memory cells when T naïve cells (T_N) encounter antigens specific to their T cell receptor (TCR) then express survival factors that permit their persistence. With the help of additional signaling and expression of anti-apoptotic factors, differentiation into several different memory phenotypes can occur, representing a spectrum of proliferation, differentiation and memory potential [93]. The most differentiated memory cells, T effector memory (T_{EM}) have the highest effector function, but lowest memory potential. In contrast, the least differentiated memory cells, T stem memory (T_{SM}), have the highest capability for self-renewal, multipotent differentiation and long-term proliferation. T_{SM} pathways are shared with hematopoietic stem cells that promote self-renewal and multipotency, in addition to T cell differentiation [61, 230].

There is increasing evidence that persistence of T_{SM} populations among T cellular therapy products is associated with superior cancer therapeutic outcomes [62]. As such, considerable efforts have been expended in the isolation, ex vivo expansion, and post-infusional monitoring for T_{SM} cells. While T_{SM} cells can be considered a desirable target T cell population for therapeutics, most reside in the lymphoid tissues with only 0.2-0.3% entering the peripheral bloodstream, which makes them impractical to isolate from a single donor as a stand-alone therapeutic product [60]. Whereas ex vivo conditions have been developed to facilitate expansion and retention of differentiated T cell populations, purified peripheral blood T_{SM} cells placed in cell culture are prone to rapidly lose their stem memory-like phenotype [237, 64]. To minimize differentiation, ex vivo protocols now incorporate cytokines (IL-15; IL-21) [93, 6], Akt inhibition, antioxidants and metabolic programming to oxidative phosphorylation [147, 175]. Yet the longer T_{SM} cells are in culture, the less they exhibit stem memory-like properties, hindering large-scale expansion. As a result of these challenges to acquiring enough T_{SM} cells and maintaining the T_{SM} phenotype, alternative methods for generating larger quantities of these cells are being sought.

Umbilical cord blood (UCB) contains immature fetal hematopoietic cells (HSC). Even more undifferentiated HSC sources, embryonic stem (ES) cells, and induced pluripotent stem cells (iPSCs), are capable of producing T progenitor cells, which can be further developed into mature T cells. Reprogramming of somatic cells into iPSCs, as discovered by Yamanaka and colleagues, allows the de-differentiation of any cell type; once formed, iPSCs have the capacity for indefinite redifferentiation into a specified cell type, rendering this platform a robust method for overcoming limited T cell subset numbers [84, 154, 210].

Because the duration of anti-tumor responses has most closely correlated with T_{SM} and not T_N for reprogramming followed by iPSC expansion and redifferentiation, we hypothesized that the stem cell-like properties of T_{SM} would provide a superior starting population for redifferentiation into T_{SM} than T_N cells (collectively known as T-iPSCs) due to developmental plasticity and retention of the epigenetic landscape and of iPSCs reprogrammed from T_{SM} cells. Further, we hypothesized that iPSCs reprogrammed from fibroblasts (FB-iPSCs) would be even more distinct from T_N- and T_{SM}-iPSCs, and thus would provide an even more inferior starting population for redifferentiation into T_{SM} cells.

To evaluate the potential efficiency of these three different starting populations for gen-

eration of redifferentiated T_{SM} cells, we first assessed the extent to which the epigenetic and transcriptomic memory was retained after reprogramming in each population. We hypothesized iPSCs with epigenetic and transcriptomic patterns similar to T_{SM} cells may provide us with a superior starting population for redifferentiation into T_{SM} cells. Following reprogramming, RNA-seq and ATAC-seq was used to perform a thorough transcriptomic and epigenetic comparison between 1) the cells of origin, 2) the T-iPSC populations (as a whole and as isolated subsets), and 3) the T-iPSCs and FB-iPSCs. Despite extensive cellular reprogramming, we found that T-iPSCs and FB-iPSC retain specific transcriptomic and epigenetic memory from their parent populations, yet distinctive epigenetic and transcriptomic identifiers for CD8(+) T cells subsets are lost in reprogramming for T_N-iPSC and T_{SM}-iPSCs. This maintenance of general transcriptomic and epigenetic memory in T-iPSCs suggest that leveraging these phenotypes as starting populations for redifferentiation could provide a limitless source of long-lasting T_N or T_{SM} cells for adoptive therapy in chemotherapy treated patients with low T_{SM} frequencies.

4.2 Reprogramming diminishes naïve and stem memory signatures in T_N-iPSC and T_{SM}-iPSC cell lines, relative to original populations

A stringent flow panel first described by the Restifo lab was used to separate T_N and T_{SM} cells from the peripheral blood mononuclear cells (PBMCs) of six healthy donors by flow cytometry sorting (Figure 4.1a) [60]. Following isolation of these individual T cell subsets, we transiently stimulated the cells to render them receptive to Sendai virus for Yamanaka factors reprogramming [203] (Figure 4.1a). Once T-iPSC cell lines were generated, validated (Figure 4.1b-g) and shown to be Sendai virus free, RNA-seq was performed on the original cell populations and iPSCs to assess the transcriptional differences and similarities to their original starting cell populations. DNA was also obtained for assay for transposase-accessible chromatin with high throughput sequencing (ATAC-seq) [25] that can identify conserved areas of accessible chromatin after reprogramming.

We first compared the quantity and effect sizes of differentially expressed genes between the original T cell populations (T_N vs T_{SM}), reprogrammed populations (T_N-iPSC vs T_{SM}-iPSC), as well as original and reprogrammed populations (T_N vs T_N-iPSC and T_{SM} vs T_{SM}-iPSC) (Figure 4.2a). Reprogramming had a dramatic change over the transcriptome in both T_N and T_{SM} cells, with significant differential expression in >50% of the genes captured by RNA-seq (Figure 4.2a). Consistently, we found that transcriptomic differences between the original T cell subsets were more pronounced than the differences between the reprogrammed populations (comparing the number of significantly differentially expressed genes over a range of fold change cutoffs; Figure 4.2a). While original T cell subsets had more than 3,000 differentially expressed genes (FDR < 0.05), reprogrammed subsets had less than 200. These results suggest that while original T cell subsets have distinct transcriptomic

differences, these differences are largely diminished after reprogramming.

Next, we performed an analysis of differential expression of gene sets from MSigDB to determine whether the T_N and T_{SM} cells maintained their naïve or stem memory signatures after reprogramming. While this analysis identified annotations synonymous with CD8(+) T_N phenotype among the original T_N cells as compared to the T_{SM} cells, no annotations associated with T_N phenotype was found when comparing the T_N-iPSC and T_{SM}-iPSC cell lines. Similarly, as compared to T_N cells, T_{SM} had increased expression of genes associated with CD8(+) T cell phenotypes more differentiated than T_N cells, consistent with the placement of T_N cells as more immature than T_{SM} cells. In contrast, an analysis of differentially expressed genes between T_{SM}-iPSC and T_N-iPSC cell lines had no significant annotations related to CD8(+) naïve or stem memory phenotypes. These results suggest that of the few genes that maintained differential expression between T_{SM}-iPSC and T_N-iPSC cell lines, these genes did not constitute annotations corresponding to CD8(+) TN or TSM phenotypes.

We next identified regions of differential chromatin accessibility between the original T_N and T_{SM} cells and between the T_N-iPSC and T_{SM}-iPSC cells lines after reprogramming. Reprogramming had a similar dramatic change over the accessible genome in both T_N and T_{SM} cells (Figure 4.2b). Although the original T_N and T_{SM} populations had marked epigenetic differences, these differences were almost entirely diminished in the reprogrammed populations (using a range of cutoff values; Figure 4.2b), consistent with the transcriptome analysis. Together, the lack of differences between the T_N and T_{SM} cells in the transcriptome and accessible regions after reprogramming suggests that induced pluripotency eliminates many of the original developmentally acquired differences observed between the T_N and T_{SM} cells.

4.3 Reprogrammed CD8(+) T cells gain pluripotent signatures while losing T cell signatures

To better understand the phenotypic differences between the original and reprogrammed T cells, we evaluated transcriptional and epigenetic changes that were maintained in the T_N-iPSCs and T_{SM}-iPSC cell lines, relative to their starting T cell populations. First, we observed that overall changes in gene expression after reprogramming were highly consistent between the two T cell populations and did not correlate with their original populations (Figure 4.2d, Figure 4.2c). To better understand these changes, we next analyzed genes that were significantly differentially expressed between the T_N-iPSCs and T_{SM}-iPSCs and their original T cell populations. While the T_N-iPSCs and T_{SM}-iPSCs both showed preferential expression of pluripotent and embryonic stem cell associated genes (SOX2/9/11, NANOG, LIN28A/B), genes related to T cell activation, including KLRK1 (NKG2D), CD62L/SELL, and all 4 peptides that form CD3 (CD3 g,d,e) and CD8$\alpha\beta$, were more highly expressed in the original T_N and T_{SM} cell populations (Figure 4.2f) [226]. While transcriptional differences between the original T_N and T_{SM} cells were observed in genes associated with naïve and

stem memory phenotype, the T_N-iPSCs and T_{SM}-iPSC cell lines did not maintain such transcriptional differences after reprogramming (Figure 4.2d) [60, 146, 153, 188]. These results show that while pluripotent signatures are acquired after reprogramming, naïve and stem memory phenotype is lost in T_N-iPSCs and T_{SM}-iPSC cell lines, respectively.

We next sought to determine whether regions near to T cell associated genes retained accessibility after reprogramming. We evaluated changes in accessibility between original and reprogrammed cells in regions proximal to genes related to pluripotency and naïve and stem memory T cell phenotypes. While accessibility around pluripotent associated genes was increased in all T-iPSC cell lines, genes related to naïve and stem memory T cell phenotypes lost accessibility after reprogramming (Figure 4.2e). Figure 4.2g shows ATAC-seq normalized read counts across the four populations surrounding CD62L/SELL and CD48, which have increased accessibility at promoters in both the T_N and T_{SM} cell types. Figure 4.2h shows ATAC-seq normalized read counts surrounding pluripotent genes SOX9 and SOX2, which showed increased accessibility in the T-iPSCs at promoters [90]. Together, changes in accessibility and expression after reprogramming are consistent with our conclusion that key naïve and stem memory T cell signatures were lost and pluripotent signatures were gained.

4.4 Transcriptomic and epigenetic memory specific to naïve or stem memory T cell phenotypes are retained after reprogramming of T_N and T_{SM} isolated T cell subsets

We next evaluated whether any transcriptomic and epigenetic differences between the T_N and T_{SM} cells were maintained after reprogramming. Figure 4.2c shows that genes that were significantly differentially expressed between T_N-iPSC and T_{SM}-iPSC cell lines did not show similar patterns of fold change expression between T_N and T_{SM} cells, respectively. To evaluate maintenance of the transcriptome, we selected all genes that were differentially expressed between the original T_N and T_{SM} cells both before and after reprogramming (Figure 4.3a). As noted above, reprogramming largely diminished the naïve and stem memory T cell specific signatures. However, 41 genes retained differential expression between T_N and T_{SM} cells after reprogramming. Of these, only 23 were differentially expressed between the T_N-iPSC and T_{SM}-iPSC cell lines in the same direction of their parent populations. Among these 23 genes was ZBTB7B, which is involved in CD8$\alpha\beta$ T cell differentiation [224] as well as other immune related genes including APOBEC3B [201], BCL3 [32], and FAS/CD95. While FAS/CD95 was not expressed in T_N cells, it plays an increasingly important role through and up to the development of effector memory CD8 T cells [60].

By selecting all regions that maintained differential accessibility between the T_N and T_{SM} cells both before and after reprogramming, we identified only 6 genomic regions that retained

accessibility (Figure 4.3b). A notable example is the region near ORM1, an acute-phase reactant involved in sphingolipid metabolism, crucial to T cell signaling (Figure 4.3b) [21, 122]. These results suggest that, although rare, memory of the original naïve and stem memory T cell phenotypes remain after reprogramming.

4.5 Reprogrammed fibroblasts lose fibroblast specific annotations identified in parent populations

Because the T_N-iPSCs and T_{SM}-iPSCs lose a majority of the naïve and stem memory phenotype seen in original populations, we next sought to compare the T-iPSCs to reprogrammed cells derived from a starting population vastly different from T cells, FB-iPSCs. We hypothesized that although the T_N and T_{SM} cells lost a majority of naïve and stem memory phenotype after reprogramming, more dissimilar cell types, such as T cells and fibroblasts, may maintain epigenetic and transcriptomic memory when compared to each other. Fibroblasts from three healthy donors were reprogrammed into FB-iPSCs (Figure 4.4a). After validating iPSCs as performed for T-iPSCs (Figure 4.4b-e), RNA was obtained for ATAC-seq and RNA-seq and processed as specified in Section 4.10.5. We next identified differentially expressed genes and differentially accessible regions between FB-iPSCs and fully differentiated fibroblasts. Because the original fibroblasts all had been reprogrammed in FB-iPSCs, we leveraged ATAC-seq from HSV-1 infected primary human fibroblasts at 0hr time points (2 replicates, GEO accession GSE100611) [74] and PolyA RNA-seq from skin fibroblasts (ENCODE, accession ENCSR510QZW) [34, 35] that are considered comparable to the original fibroblast starting populations. Upon initial transcriptome comparison between FB-iPSCs and fibroblasts, over half of the genes captured by RNA-seq were differential between fibroblasts and FB-iPSCs (Figure 4.5a, FB vs FB-iPSC). Similarly, accessible regions between the FB-iPSCs and fibroblasts were overridden (Figure 4.5b, FB vs FB-iPSC). Thus, both transcriptome and epigenome analyses demonstrated drastic changes in FB-iPSCs compared to fibroblasts, and of similar magnitude to that observed after reprogramming in T cells.

We next sought to understand which genes were differentially expressed between fibroblasts before and after reprogramming. Figure 4.5c shows log2 fold change between FB-iPSCs and fibroblasts and corresponding adjusted p-values for all genes. Pluripotent associated genes such as NANOG, LIN28A/B, and SOX2 had increased expression in FB-iPSCs, whereas collagen associated genes and fibroblast growth factors, components of the extracellular matrix and known to be synthesized by fibroblasts, were had increased expression in the original starting population [150]. Figure 4.5d shows corresponding changes in accessibility proximal to key pluripotent related and fibroblast related genes. From MSigDB annotated, genes associated with stemness and pluripotency are shown in blue, and those associated with collagen, fibroblast growth, and fibroblast proliferation are yellow. Many pluripotent associated genes had increased expression and corresponding chromatin accessibility after reprogramming, in contrast to decreased fibroblast related genes expression and chromatin

accessibility in FB-iPSCs. Changes in chromatin accessibility are depicted in Figure 4.5e that shows increased NANOG promoter and decreased collagen associated gene COL3A1 accessibility in FB-iPSCs, paralleling the loss of accessibility near T cell associated genes after T_N and T_{SM} cell reprogramming.

4.6 Reprogrammed fibroblasts and T cells retain epigenetic and transcriptomic memory of original starting populations

We next compared changes in the transcriptome and epigenome between reprogrammed T cells and fibroblasts to determine whether memory was maintained between original starting populations. Because the T_N-iPSCs and T_{SM}-iPSCs had similar transcriptomes and patterns of accessibility, we grouped both the T_N-iPSCs and T_{SM}-iPSCs as T-iPSCs in our comparison to the FB-iPSCs. We first identified differentially expressed genes and differentially accessible regions between the T-iPSCs and FB-iPSCs and found that differences between the T_N-iPSCs and T_{SM}-iPSCs were largely diminished after reprogramming (Figures 4.2a and 4.2b). In contrast, the FB-iPSCs and T-iPSCs maintained significant differences, as evidenced by 4,755 significantly differentially expressed genes and 28,546 differentially accessible regions between reprogrammed populations (Figures 4.5a and 4.5b). We next evaluated the overlap between genes that were significantly differentially expressed between T cells and fibroblasts in both original and reprogrammed populations to evaluate whether genes associated with fibroblast or T cell phenotypes were retained after reprogramming. We found significant overlap in the genes and accessible regions that were maintained after reprogramming (Figure 4.5f). These results suggest that both accessibility and transcriptomic memory is retained in T-iPSC and FB-iPSC cell lines.

We next performed an analysis to measure the extent to which gene sets, available in MSigDB, associated with fibroblast and CD8(+) T cell phenotypes retained differential expression between FB-iPSCs and T-iPSCs. Through this analysis, we identified gene sets associated with the extracellular matrix and collagen formation that had significantly increased expression in FB-iPSCs. However, using this approach, no gene sets associated with CD8(+) T cell phenotypes had significantly increased expression in the T-iPSCs.

Although gene sets associated with CD8(+) T cell phenotype did not maintain increased expression in T-iPSCs as a whole, we next determined whether individual genes associated with CD8(+) T cell phenotype maintained increased expression after reprogramming. To determine whether genes associated with CD8(+) T cell phenotype maintained increased expression in the T-iPSCs, we first computed a set of "T cell expressed genes" and "fibroblast expressed genes" from original starting populations and assessed whether these genes maintained differential expression after reprogramming. Figure 4.5g shows row normalized expression counts for a subset of genes that maintained differential expression after reprogramming. Figure 4.5g additionally compares the transcriptome of T-iPSCs and FB-

iPSCs to that of embryonic stem cells (ES) (GEO accession GSE115046) [83] to determine the extent to which these populations were reprogrammed. Although CD62L lost relative promoter accessibility and expression after reprogramming (Figures 4.2g and 4.2d), select T-iPSC clones maintained increased expression of CD62L, compared to the FB-iPSCs (Figure 4.5h). Increased expression of CD3E, a member of T-cell receptor-CD3 complex, was maintained in some T-iPSC clones [14]. CD8 associated genes maintained variable levels of expression in the T-iPSCs across clones, in contrast to FB-iPSCs and embryonic stem (ES) cells that maintained consistently lower levels of expression (Figure 4.5g).

Through analysis of differentially expressed genes between the FB-iPSCs and T-iPSCs using QIAGEN Ingenuity Pathway Analysis IPA (IPA) (QIAGEN Inc.), we found genes associated with WNT (Gordon and Nusse, 2006) and TEC Kinase [197] signaling to have increased expression in the T-iPSCs (Figure 4.5g). TEC kinase signaling is important to the development and activation of B cells and T cells. WNT signaling is of particular interest due to its known role in regulating stem cell pluripotency, consistent with the up-regulation of pluripotent associated genes SOX21/11, and NANOG in T-iPSCs. Additionally, in a principal component analysis (PCA) of normalized gene expression counts for the FB-iPSCs, T-iPSCs, and ES cells, the T-iPSCs are more similar to ES than FB-iPSCs, driven by the 2nd principal component (PC) (Figure 4.6a). PCA of ATAC-seq shows similar results, where the T-iPSCs were more similar to ES through the 2nd PC, although with a much smaller effect size (Figure 4.6b). Together, these data suggest the existence of a gradient of reprogramming, where FB-iPSCs have retained more transcriptomic and epigenetic memory from their initial starting population than T-iPSCs.

We next leveraged ATAC-seq to evaluate changes in chromatin accessibility between the T-iPSCs and FB-iPSCs nearby genes that were differentially expressed between the T-iPSCs and FB-iPSCs. Figure 4.4h shows aggregated log2 fold change in both ATAC-seq and RNA-seq in the FB-iPSCs and T-iPSCs. Here, the T-iPSCs show increased expression and accessibility near T cell expressed genes, while the FB-iPSCs maintain expression and accessibility near fibroblast related and collagen/ECM genes. We next used GREAT [134] to annotate differentially accessible regions between T-iPSCs and FB-iPSCs, and found that the T-iPSCs had increased accessibilitfy near genes associated with the alpha-beta T cell receptor complex (GO term 0042105). Figure 4.4i shows changes in accessibility surrounding CD3E and CD3D, two genes encoding components of the T-cell receptor/CD3 complex. These results suggest that although fine grained memory dictating differences between naïve and stem memory populations were lost after reprogramming, differences in gene expression and accessibility that dictate differences between CD8(+) T cells and fibroblasts were still be observed after reprogramming.

4.7 T-iPSC and FB-iPSC are differentiated into hematopoietic CD34+ cells at a similar efficiency

iPSCs were subjected to embryoid body (EB) based differentiation to obtain CD34(+) cells also present on human HSCs that can be differentiated into all blood cell lineages [13]. To generate CD34(+) cells, we guided iPSC cell differentiation into mesoderm using growth factors BMP4 and bFGF during the first two days of EB differentiation (Figure 4.7a) followed by modulating Wnt and TGF beta signaling pathways using CHIR99021 and SB431542 small molecules, respectively, prompting development towards definitive hematopoietic cells. After 9 days in a serum free, stroma free EB differentiation method, T_N/T_{SM} -iPSCs and FB-iPSCs were differentiated into CD34(+)CD43(-) hemangioendothelial cells capable of further differentiation into definitive hematopoietic cells. Despite retention of epigenetic and transcriptional differences between T-iPSCs and FB-iPSCs the frequencies of CD34(+)CD43(-) cells were similar (Figure 4.5b, Figure 4.7c).

4.8 Hematopoietic CD34(+) cell differentiation into T progenitor cells

On EB day 9, CD34(+) cells were co-cultured on a mouse stromal cell line (OP9) expressing Notch ligand DLL4 (OP9-DLL4) that is essential for robust T cell development. Cells were passaged every 3-4 days and their phenotype were analyzed using flow cytometry for T progenitor cell surface markers at indicated time points (Figure 4.7d). Figure 4.7e shows acquisition of CD3 and TCR$\alpha\beta$ as well as CD8 and CD4 double positive expression in T-iPSC derived T progenitors compared to FB-iPSC derived T progenitors over the course of 3 weeks in T specification conditions. While T progenitor cultures derived from T-iPSC lines were able to be carried out to 4 weeks in T specification culture and continued to show further T cell development, all FB-iPSC derived cultures contracted and failed to produce T progenitors of sufficient number or phenotype for further study. Additionally, Together, these preliminary results suggest that the T-iPSCs can successfully differentiate to CD34(+) hematopoietic lineage cells that upon four weeks of differentiation will acquire key T progenitor cell surface markers that will form mature T cells in a culture system designed for T cell maturation.

4.9 Discussion

Overall, this analysis of T-iPSCs and FB-iPSCs demonstrates epigenetic and transcriptomic maintenance of memory after reprogramming, although distinct T_N and T_{SM} phenotypes were lost. Despite this maintenance of memory, preliminary results show that similar frequencies of hematopoietic CD34(+) cells were obtained from the T-iPSCs and FB-iPSCs

after 9 days of differentiation. These results highlight the potential importance of considering starting populations for reprogramming and redifferentiation to T cells, and the need for improved differentiation systems to evaluate the effect of epigenetic maintenance on redifferentiated populations.

Although most somatic cell types have been shown to successfully reprogram to iPSCs [71, 159, 203], considering the starting population of iPSCs is crucial to the efficacy of producing T lymphocytes. Different tissues demonstrate variable amenability to reprogramming into iPSCs [7, 129]. Studies show that iPSCs can maintain memory of their tissues of origin [46, 104, 155]. Kim et al. reported that DNA methylation patterns of blood derived and fibroblast derived iPSCs maintained residual methylation patterns to their tissue of origin [104]. Efrat et al. showed significant differences in patterns of chromatin accessibility between human pancreatic islet β cells and FB- iPSCs that are maintained from their tissues of origin [46]. Here, we similarly found that T cells derived from T_N-iPSCs and T_{SM}-iPSCs did not retain epigenetic memory unique to their starting population. Additionally, we found that the transcriptomics and chromatin accessibility of the T-iPSCs are more similar to ES cells than FB-iPSCs, consistent with previous studies that show more prominent differences in epigenetics between FB-iPSCs and ES cells, as compared to blood derived iPSCs [104].

While the reprogramming efficacy of fibroblasts has been estimated to be 10-50 times higher than T cells, it remains unclear if reprogrammed T-iPSC clones have an advantage over FB-iPSC clones during differentiation to T cells as a result of retained T cell specific priming [199]. Vizcardo et al. demonstrated that CD8(+) T cells derived iPSCs yield the highest proportion of TCRβ(+)CD3(+) cells, when compared to iPSCs derived from CD34(+) cord blood cells and CD3(+) T cells [222]. Our preliminary redifferentiation results show that despite retention of epigenetic and transcriptional differences between T-iPSCs and FB-iPSCs, similar frequencies of hematopoietic CD34(+)CD43(-) cells that give rise to Tprogenitor cells was observed from TN-, TSM-, and FB-iPSCs. Nonetheless, redifferentiating iPSCs to CD8(+)CD4(-)T cells remains challenging. Currently, protocols for differentiation of T cells from iPSCs suffer from poor reproducibility especially for different clones and provide inconsistent cell yields of CD8(+)CD4(+) T cells, which limits scaled production for clinical translation [149]. Further, most differentiation protocols, including the one employed by our group, utilizes stromal feeder cells and serum, which can add culture-to-culture variation and are generally viewed as impractical for clinical trials [154]. A recent study published by Iriguchi et al. detailed their method for T-iPSC differentiation, which demonstrated reproducible and scalable generation of functional CD8ab(+) T cells without using stromal feeder cells in any stage of the differentiation process [84]. To accomplish this, SDF1a and a p38 inhibitor were added to the culture to more closely mimic thymopoiesis, increasing yields of CD8(+)CD4(+)T cells that remained highly amenable to maturation and expansion [84]. For this purpose, serum was required that could be problematic for clinical translation. As a large fraction of the end product T cells switched CD45 isotype from CD45RA (naïve) to CD45RO (memory) during the maturation and expansion process, consistent with more differentiated cells that may have a shorter longevity upon in vivo adoptive transfer. Further protocol refinements are ongoing. A collaboration between Dr.

Sadelin and Fate Therapeutics to produce anti-cancer T cells from iPSCs is imminent for translation into the clinic and results from such a trial will be early awaited.

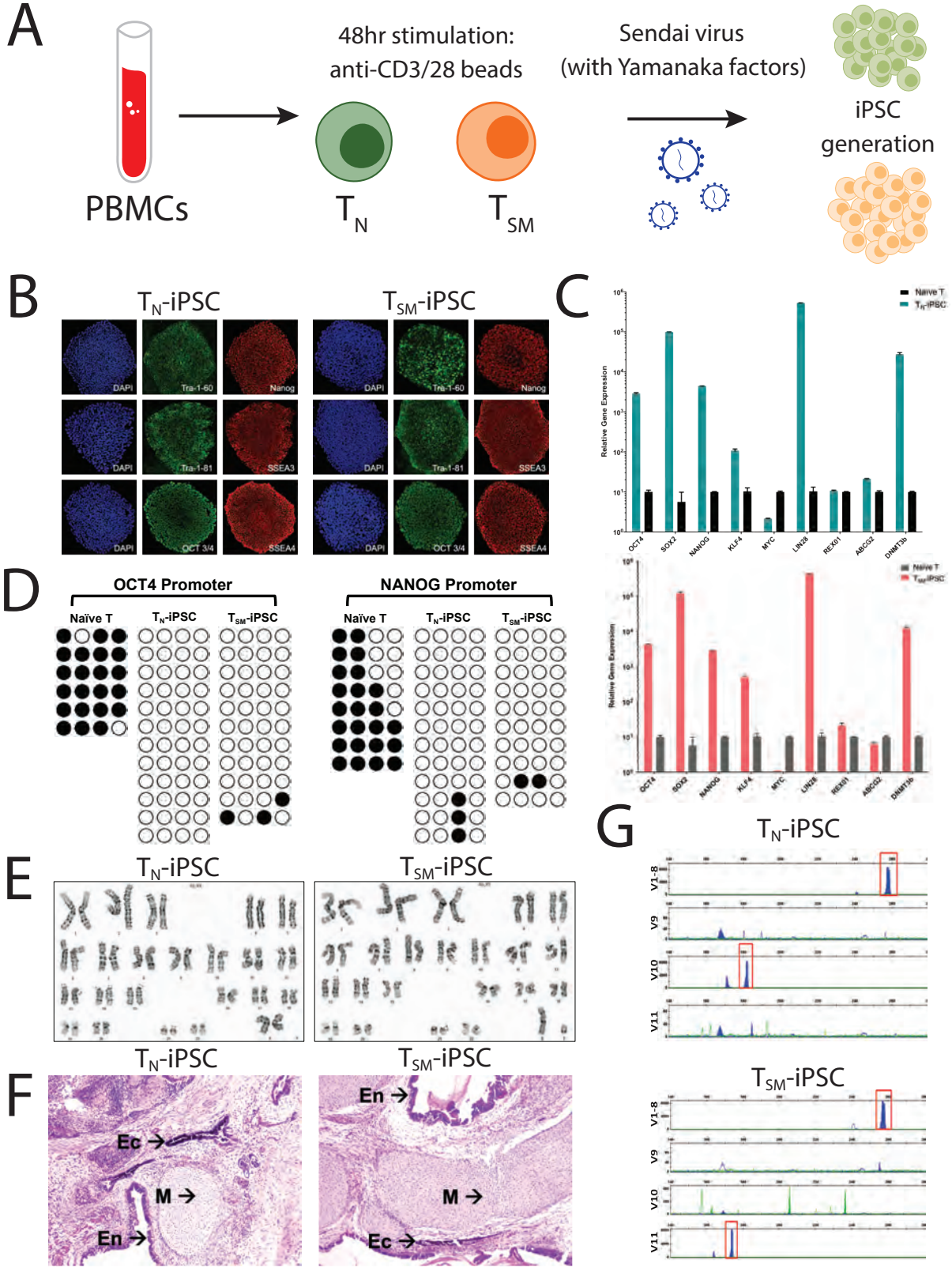

A
PBMCs
48hr stimulation:
anti-CD3/28 beads
T N
T SM
Sendai virus
(with Yamanaka factors)
iPSC
generation
B
T N-iPSC
T SM-iPSC
DAPI
Tra-1-60
Nanog
Tra-1-81
SSEA3
OCT 3/4
SSEA4
C
Naïve T
T N-iPSC
Relative Gene Expression
OCT4
SOX2
NANOG
KLF4
MYC
LIN28
REX01
ABCG2
DNMT3b
D
OCT4 Promoter
NANOG Promoter
Naïve T
T N-iPSC
T SM-iPSC
E
T N-iPSC
T SM-iPSC
F
T N-iPSC
T SM-iPSC
Ec
M
En
G
T N-iPSC
T SM-iPSC
V1-8
V9
V10
V11

Figure 4.1

Figure 4.1: T_N and T_{SM} cells were reprogrammed from 3 donors each to T_N-iPSC and T_{SM}-iPSC and characterized for pluripotency. a) PBMCs were extracted from six healthy donors, flow sorted based on strict criteria for differentiating T_N and T_{SM} cells (3 donors each), and then incubated with Sendai virus to generate T cell derived induced pluripotent stem cells (T-iPSCs). RNA-seq was collected for T_N (3 donors, 1-2 replicates) and T_{SM} (3 donors, 1-2 replicates), T_N-iPSC (3 donors, 4 replicates), and T_{SM}-iPSC (3 donors, 4 replicates). ATAC-seq was collected for T_N (3 donors, 1 replicate) and T_{SM} (3 donors, 1 replicate), T_N-iPSC (3 donors, 2 replicates), and T_{SM}-iPSC (3 donors, 2 replicates). b) Immunofluorescence staining for pluripotency markers TRA-1-60, NANOG, TRA-1-81, SSEA3, OCT-3/4, and SSEA4 in T_{SM} and T_N-iPSC colonies. c) Quantitative RT-PCR based analysis of pluripotency gene expression in T-iPSC clones. Parental T naïve (T_N) cells used as a negative control. d) Bisulfite sequencing analysis demonstrating CpG hypomethylation at Nanog and Oct4 promoter regions in T-iPSC clones. e) Chromosome spread demonstrating that T-iPSC clones possess a normal karyotype. f) Teratoma formation assay shows T-iPSC clones formed all three germ layers Ec (ectoderm), M (mesoderm) and En (endoderm) labeled in the images. g) T-cell receptor rearrangement analysis demonstrating the presence of unique, clonal TCR rearrangements in the T-iPSC clones.

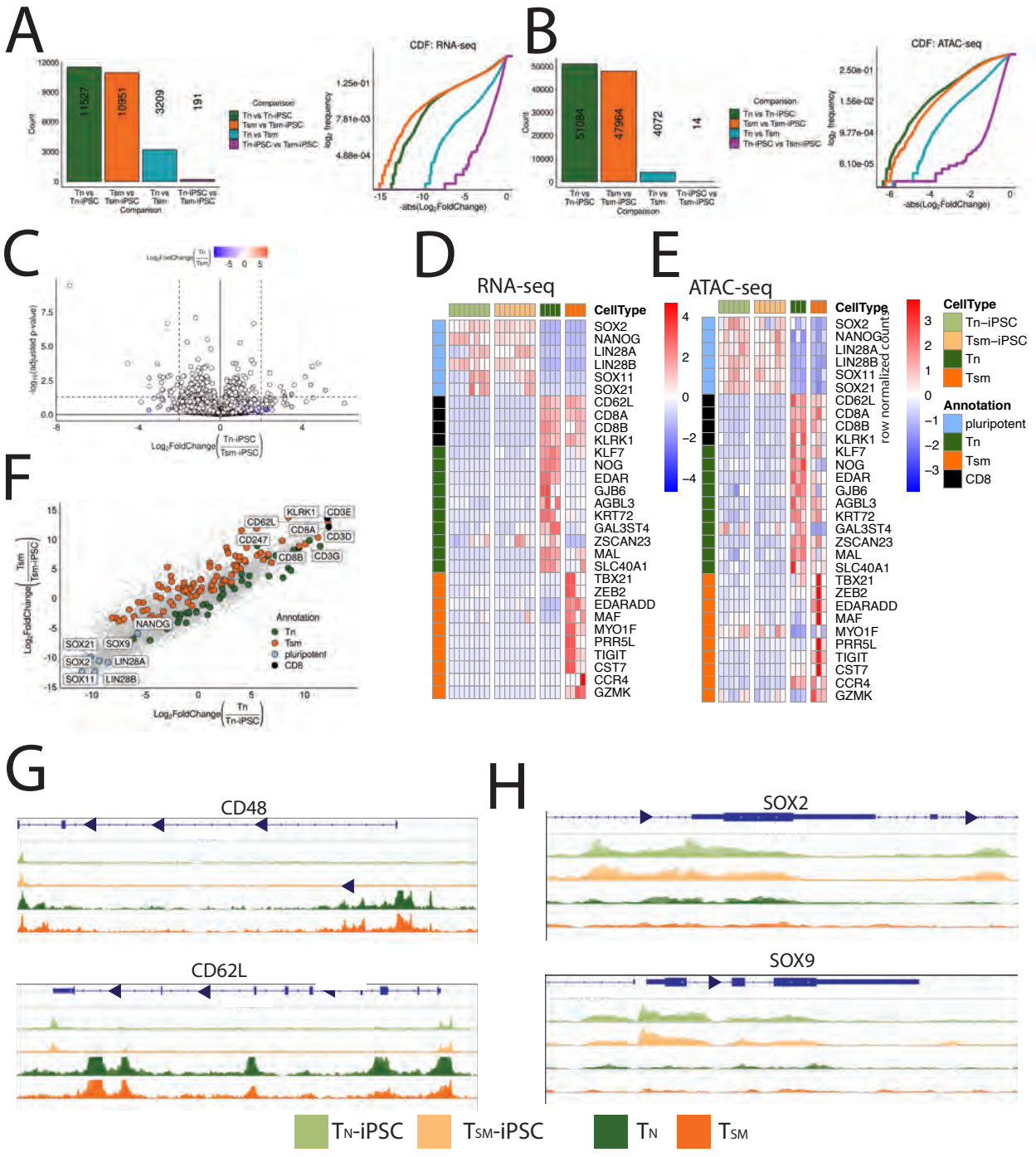
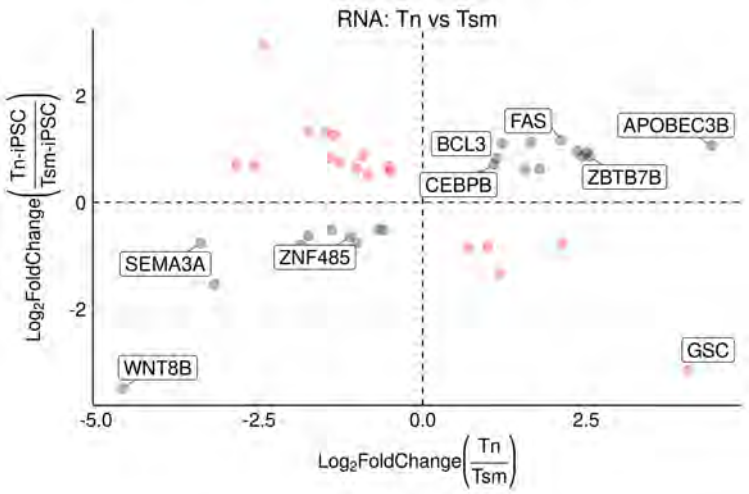
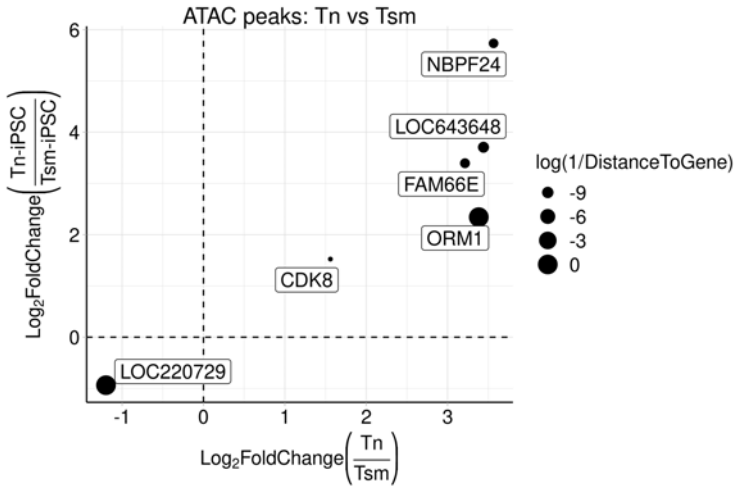

A
B
C
D
E
F
G
H
CDF: RNA-seq
CDF: ATAC-seq
RNA-seq
ATAC-seq
CellType
Annotation
CD48
CD62L
SOX2
SOX9
TN-iPSC
TSM-iPSC
TN
TSM

Figure 4.2

Figure 4.2: T_N and T_{SM} cells were reprogrammed from 3 donors each to T_N-iPSC and T_{SM}-iPSC and characterized for pluripotency. a) (Left) Counts of genes that were significantly differentially expressed (significance defined as abs(log2foldchange) > 0.5 and adjusted p-value < 0.05) between the following four conditions: T_N vs T_N-iPSC, T_{SM} vs T_{SM}-iPSC, T_N vs T_{SM}, and T_N-iPSC vs T_{SM}-iPSC. (Right) Cumulative distribution functions (CDFs) of log2 fold change of differentially expressed genes between each of the four conditions. b) (Left) Counts of regions that were significantly differentially accessible (significance defined as abs(log2foldchange) > 0.5 and adjusted p-value < 0.05) between the following four conditions: T_N vs T_N-iPSC, T_{SM} vs T_{SM}-iPSC, T_N vs T_{SM}, and T_N-iPSC vs T_{SM}-iPSC. (Right) Cumulative distribution functions (CDFs) of log2 fold change of differentially accessible regions between each of the four conditions. c) Volcano plot of log2 fold change in expression between T_N-iPSCs and T_{SM}-iPSCs (x-axis) and -log10 adjusted p-values (y-axis). Results are colored by log2 fold change expression between the original CD8 T_N and T_{SM} subsets. d) Row normalized expression of T_N-iPSC, T_{SM}-iPSC, T_N and T_{SM} cells for pluripotent, CD8 T cell, and naïve and stem memory T cell subset related genes. e) Mean counts of ATAC-seq cut sites overlapping peaks near pluripotent, CD8 T cell, and naïve and stem memory T cell subset related genes. Counts from multiple ATAC-seq peaks nearest to a gene were combined by computing mean cut sites. Results are row normalized. f) Log2 fold change in gene expression between T_N-iPSC cell lines and the original T_N cells (x-axis) and T_{SM}-iPSC cell lines and the original T_{SM} cells (y-axis). CD8 T cell and pluripotent associated genes are labeled. Negative numbers indicate increased expression in T-iPSC cell lines, while positive numbers indicate increased expression in the original T cells. g) Aggregated read counts of ATAC-seq at SELL and CD48 loci show diminished accessibility around promoter regions of reprogrammed CD8 T cells. h) Aggregated read counts of ATAC-seq at Sox9 and Sox2 show increased accessibility around promoter regions of reprogrammed CD8 T cells.



(a)



(b)

Figure 4.3: Shared accessible peaks and genes maintained after reprogramming T_N-iPSCs and T_{SM}-iPSCs from starting CD8(+) T cell populations. (a) Log2 fold change expression of significantly differentially expressed genes (FDR < 0.05) between T_N and T_{SM} (x-axis) and T_N-iPSCs and T_{SM}-iPSCs (y-axis). Red dots indicate genes that are differentially expressed but do not have correlated directionality between the T-iPSC and original T cell populations. Black dots indicate genes that have correlated directionality between the T-iPSC and original T cell populations. (b) Log2 fold change in accessibility of significantly differentially accessible regions (adjusted p-value < 0.05) between T_N and T_{SM} (x-axis) and T_N-iPSCs and T_{SM}-iPSCs (y-axis).

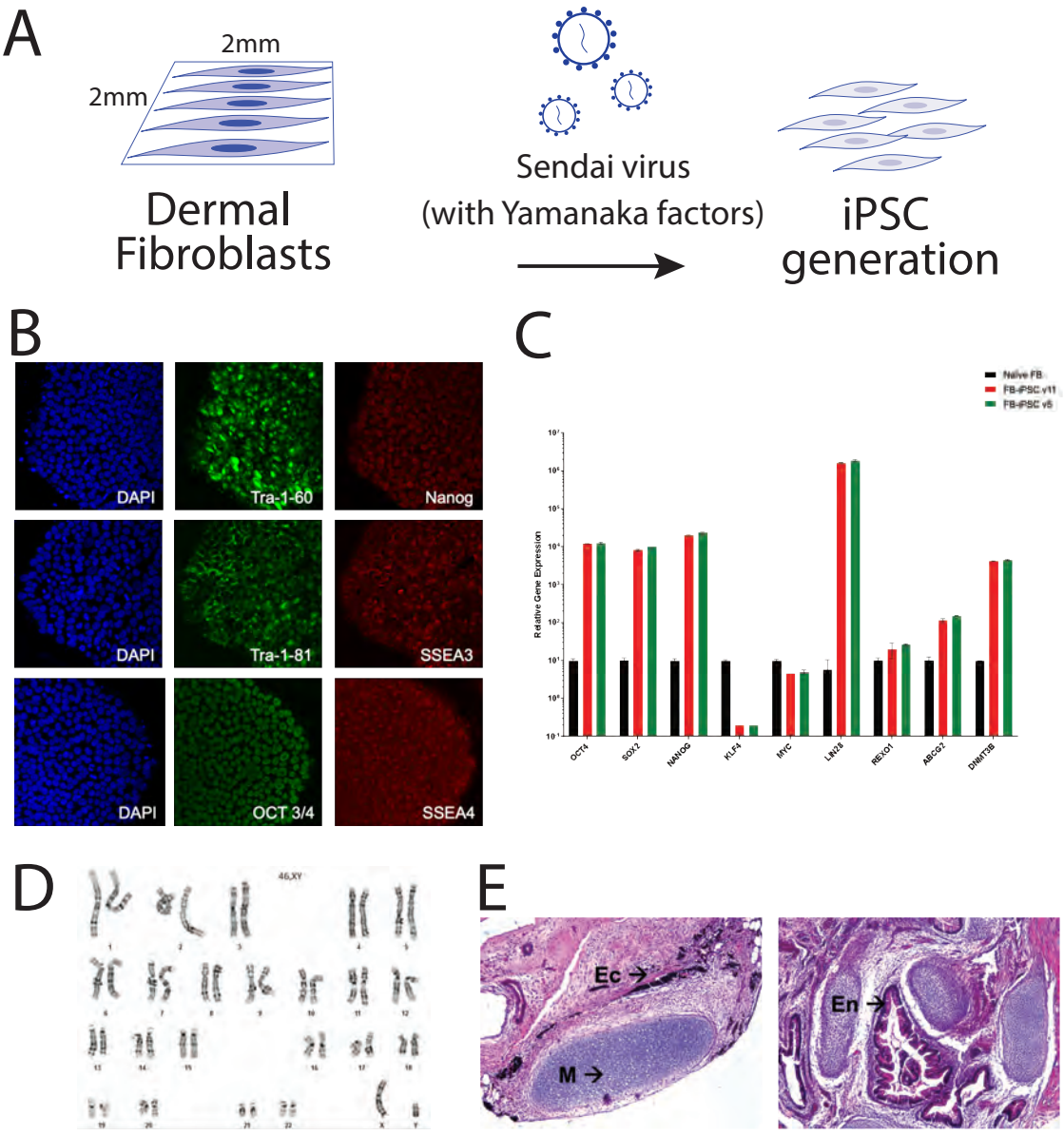

A
2mm
2mm
Dermal Fibroblasts
Sendai virus
(with Yamanaka factors)
iPSC generation
B
DAPI
Tra-1-60
Nanog
DAPI
Tra-1-81
SSEA3
DAPI
OCT 3/4
SSEA4
C
Relative Gene Expression
OCT4
SOX2
NANOG
KLF4
MYC
LIN28
REXO1
ABCG2
DNMT3B
D
46,XY
E
Ec
M
En

Figure 4.4

Figure 4.4: Fibroblasts were reprogrammed from three donors to FB-iPSCs and characterized for pluripotency. a) Dermal fibroblasts (FB) were isolated from skin biopsies and cultured in human FB medium. FB cells were reprogrammed using Sendai virus. FBs were infected with four retroviral supernatants (Yamanaka factors OCT4, SOX2, KLF4, and c-MYC) for reprogramming to FB-iPSCs. RNA-seq was collected for FB-iPSCs (3 donors, 4 replicates). RNA-seq for fibroblasts were taken from ENCODE (accession ENCSR510QZW) (2 replicates). ATAC-seq was collected for FB-iPSCs (3 donors, 2 replicates). ATAC-seq for fibroblasts were taken from GEO (accession GSE100611, 2 biological replicates). b) Immunofluorescence staining for pluripotency markers TRA-1-60, NANOG, TRA-1-81, SSEA3, OCT-3/4, and SSEA4 in FB-iPSC clone 5. c) Quantitative RT-PCR based analysis of pluripotency gene expression in FB-iPSC clones. Naïve FB cells used as a negative control. d) Chromosome spread demonstrating that FB-iPSC clone 5 possess a normal karyotype. e) Teratoma formation assay shows FB-iPSC clone 5 formed all three germ layers Ec (ectoderm), M (mesoderm) and En (endoderm) labeled in the images.

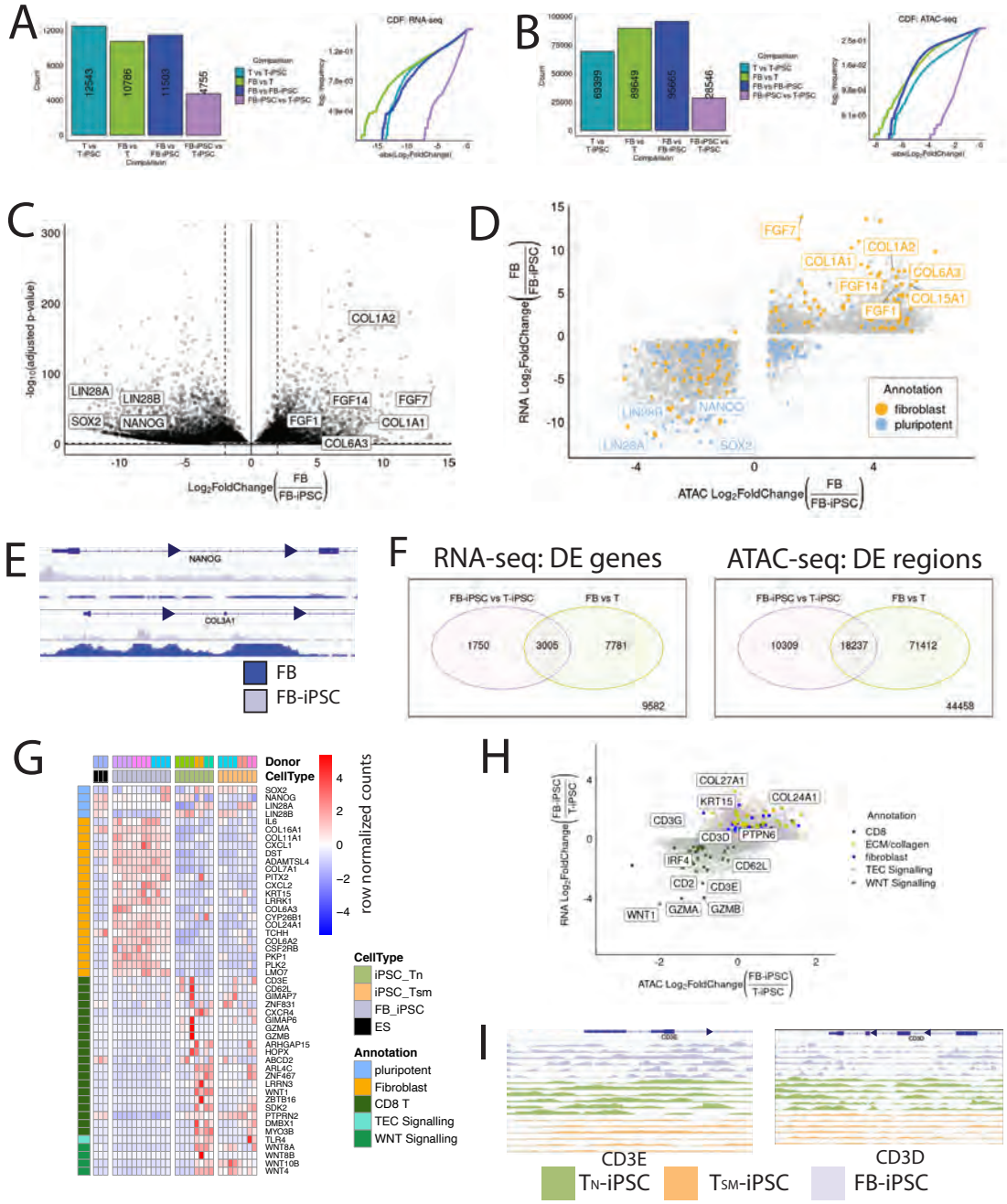
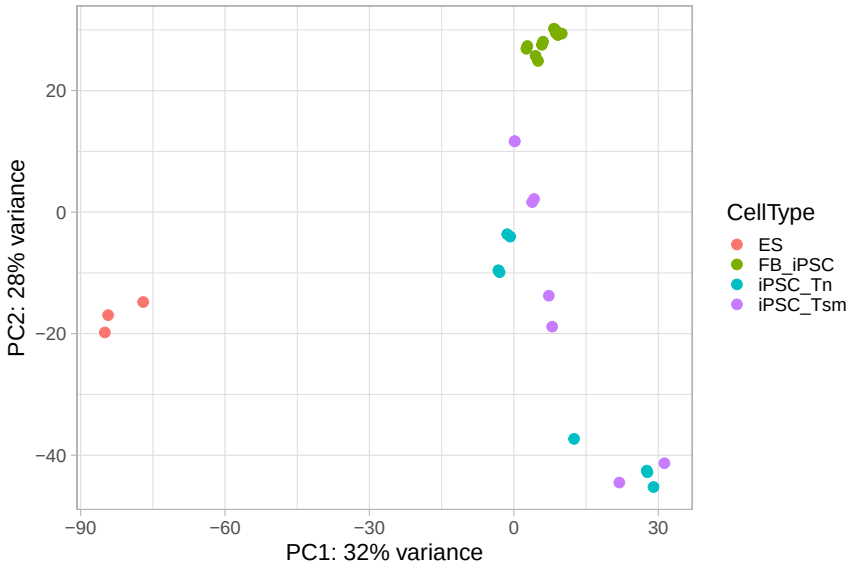
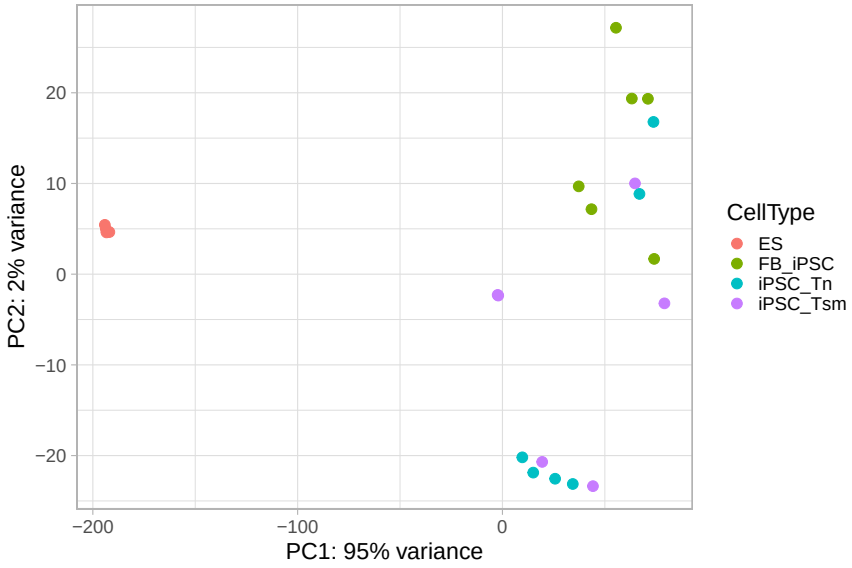

A
B
C
D
E
F
G
H
I
RNA-seq: DE genes
ATAC-seq: DE regions
FB
FB-iPSC
CD3E
CD3D
TN-iPSC
TSM-iPSC
FB-iPSC

Figure 4.5

Figure 4.5: Reprogrammed fibroblasts and T cell retain epigenetic and transcriptomic memory after reprogramming. a) (Left) Counts of genes that were significantly differentially expressed (significance defined as abs (log2foldchange) > 0.5 and adjusted p-value < 0.05) between the following four conditions: CD8(+) T cell vs T-iPSC, fibroblast (FB) vs T cell, FB vs FB-iPSC, and FB-iPSC vs T-iPSC. T cells and T-iPSC consist of T_N and T_{SM}s. (Right) Cumulative distribution functions (CDFs) of log2 fold change of differentially expressed genes between each of the four conditions. b) (Left) Counts of regions that were significantly differentially accessible (significance defined as abs(log2foldchange) > 0.5 and adjusted p-value < 0.05) between the following four conditions: T cell vs T-iPSC, fibroblast (FB) vs T cell, FB vs FB-iPSC, and FB-iPSC vs T-iPSC. (Right) Cumulative distribution functions (CDFs) of log2 fold change of differentially accessible regions between each of the four conditions. c) Volcano plot of log2 fold change and -log10 adjusted p-values of differential expression for genes compared in fibroblasts and FB-iPSCs. d) Joint analysis of log2 fold change of ATAC-seq and RNA-seq between fibroblasts (FB) and FB-iPSCs. x-axis shows ATAC-seq log2 fold change between fibroblasts and FB-iPSCs. y-axis shows RNA-seq log2 fold change. For each gene on the y-axis, the closest ATAC-seq peak to a given gene with the greatest absolute log2 fold change is shown on the x-axis. Size indicates the inverse distance to the gene of interest. Bottom left quadrant shows genes with increased expression and accessibility in fibroblasts Top right quadrant shows genes with increased expression and accessibility in FB-iPSCs. e) Example ATAC-seq normalized Tn5 cut site counts in fibroblasts and FB-iPSCs surrounding pluripotent associated gene NANOG and collagen associated gene COL3A1. f) Venn diagrams show overlap between differentially expressed genes (left) and differentially accessible regions (right) between original (T vs fibroblast) and reprogrammed (FB-iPSC vs T-iPSC) populations. Differential genes and regions are selected with FDR < 0.05 and absolute value of log2 fold change > 0.5. g) Row normalized expression of selected genes in FB-iPSCs, -iPSCs, and T_{SM}s. Selected genes include pluripotent associated, collagen/extracellular matrix associated, and CD8 associated genes. All genes displayed are differentially expressed between FB-iPSCs and T-iPSCs. Columns are sorted within each cell type by donor. h) Joint analysis of log2 fold change of ATAC-seq and RNA-seq between T-iPSC and FB-iPSC conditions. x-axis shows ATAC-seq log2 fold change between FB-iPSC and T-iPSC. y-axis shows RNA-seq log2 fold change. For each gene on the y-axis, the closest ATAC-seq peak to a given gene with the greatest absolute log2 fold change is shown on the x-axis. Size indicates the inverse distance to the gene of interest. i) Aggregated normalized ATAC-seq read counts in T-iPSCs and FB-iPSCs at genes CD3E and CD3D, associated with the α/β T cell receptor.



(a)



(b)

Figure 4.6: Principal component analysis (PCA) of ATAC-seq and RNA-seq samples shows variability in similarity of reprogrammed T cells and fibroblasts to embryonic stem cells. (a) PCA of RNA-seq of 5,000 highest variable genes. (b) PCA of ATAC-seq of top 1,000 variable peak regions.

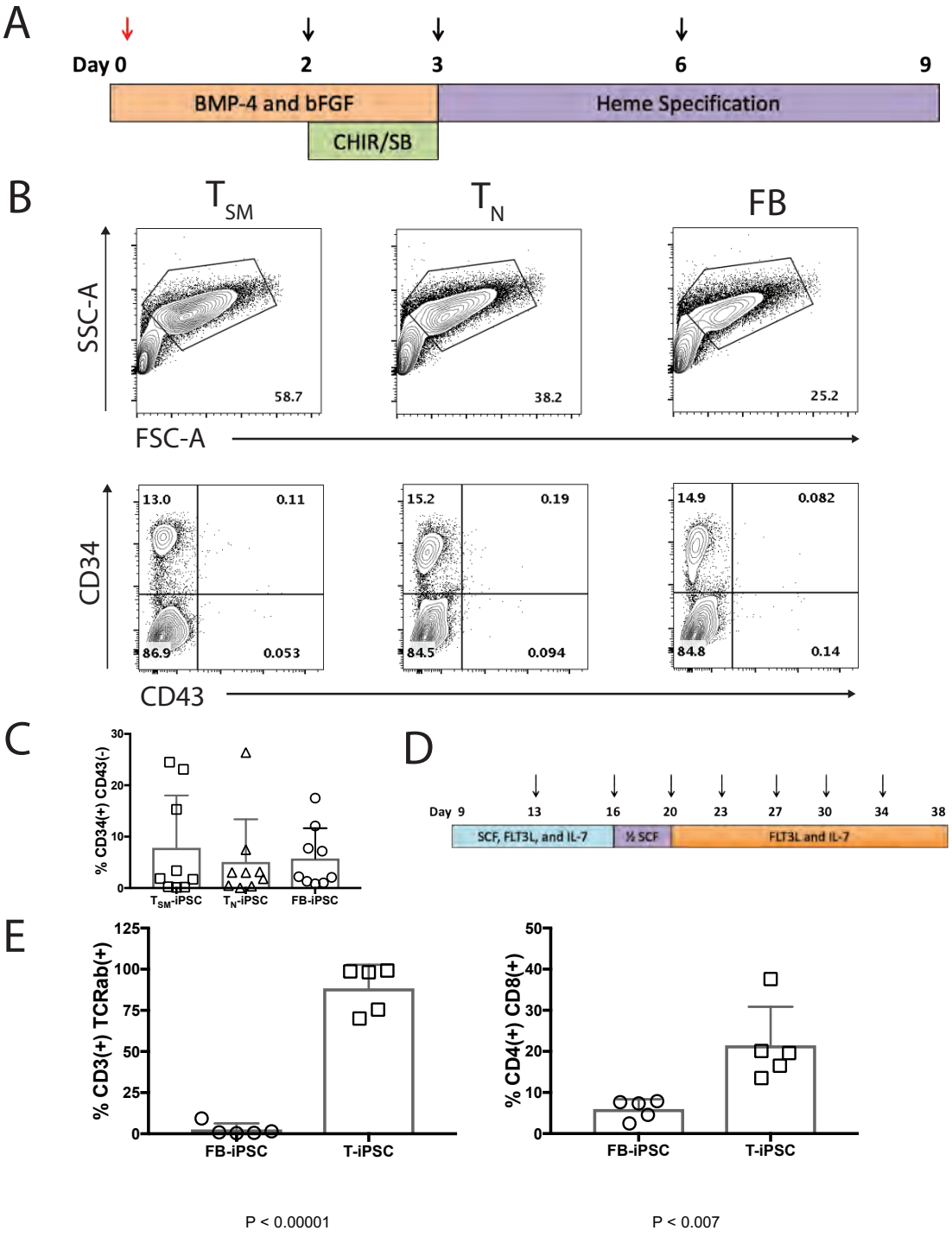

A
Day 0
2
3
6
9
BMP-4 and bFGF
CHIR/SB
Heme Specification
B
T SM
T N
FB
SSC-A
FSC-A
58.7
38.2
25.2
CD34
CD43
13.0
0.11
86.9
0.053
15.2
0.19
84.5
0.094
14.9
0.082
84.8
0.14
C
% CD34(+) CD43(-)
T SM-iPSC
T N-iPSC
FB-iPSC
D
Day 9
13
16
20
23
27
30
34
38
SCF, FLT3L, and IL-7
½ SCF
FLT3L and IL-7
E
% CD3(+) TCRab(+)
FB-iPSC
T-iPSC
P < 0.00001
% CD4(+) CD8(+)
FB-iPSC
T-iPSC
P < 0.007

Figure 4.7

Figure 4.7: T-iPSCs and FB-iPSCs derived embryoid bodies generate CD34(+) hematopoietic cells that acquire T cell surface markers after 4 weeks in culture. a) Schematic of culture conditions embryoid bodies generated using T-iPSCs or FB-iPSCs. Media changes were done on days 2, 3, and 6. Cells were harvested at day 9 of embryoid body culture. b) Flow cytometry plots showing phenotype of dissociated embryoid bodies. c) Graphic depiction of day 9 output of CD34(+) CD43(-) cells for 9 replicates of T_{SM}-iPSC, T_N-iPSC, and FB-iPSC lines. d) Schematic showing culture conditions of CD34(+) cells isolated from embryoid bodies. Isolated CD34(+) cells are cultured on plates coated with confluent OP9-DLL4 with the cytokine combinations as noted. Cells were passaged every 3-4 days as indicated by arrows, then harvested at day 38. Arrows indicate passage and flow analysis. e) Plots show acquisition of CD3 and TCR$\alpha\beta$ (left) and CD4 and CD8 (right) in FB-iPSC vs T-iPSC (3 replicates from T_N-iPSC line and 2 replicates from T_{SM}-iPSC line) derived T progenitors over 3 weeks in T specification culture.

4.10 Methods and Materials

All experimental methods discussed in this Section were performed by the Blazar Laboratory at the University of Minnesota.

4.10.1 Flow sorting of T_N and T_{SM} cells from human peripheral blood

Record exact volume of apheresis unit from Memorial Blood Center. Add 5 ml Acid Citrate Dextrose Solution (ACD-A) per 40 ml apheresis unit. Transfer blood to a sterile container and add an equal volume of ammonium chloride (STEMCELL Technologies catalog# 07850) Mix by inverting gently; incubate on ice for 10 minutes. Centrifuge sample at 500 X g at room temperature (RT) 15-25°C for 10 minutes. Remove supernatant and wash cells with column buffer, spin at 150 X g for 10 minutes with break off. Carefully remove supernatant. Repeat wash step until platelets are mostly removed (1-2X). Follow the STEMCELL Technologies EasySep kit instructions for Human naïve CD8 T cell selection (Catalog#19158). This procedure is for processing 500 µl- 8 ml of sample (up to 4 X 108 cells). Prepare the MNC suspension at a concentration of 5 X107 cells/ml in recommended medium (see below). Add the EasySep Human Naïve CD8 T cell enrichment cocktail at 50 µl/ml cells. Immediately add the EasySep Human CD45RO depletion cocktail at 50 µl/ml. Mix well and incubate at RT for 30 minutes. Vortex the EasySep D2 magnetic particles for approximately 30 seconds. Ensure that the particles are in a uniform suspension with no visible aggregates. Add EasySep D2 magnetic particles at 100 µl/ml cells. Mix well and incubate at RT for 10 minutes. Bring cell suspension up to 10 ml with recommended medium. Mix cells by pipetting up and down then place tube on magnet without the cap. Set aside for 5 minutes.

Pick up magnet and in one continuous motion, invert the magnet and tube, pouring off the desired fraction into a new 14 ml tube. The unwanted magnetically labeled cells will remain bound in the original tube. Leave magnet and tube inverted for 2-3 seconds, then return to upright position. Do not shake or blot off drops that may remain hanging from the tube. Remove original tube from magnet and replace with the tube holding desired cells and set aside for 5 more minutes for a total of 2 X 5 minute separations. Cells are ready for counting and staining. Stain cells with viability dye (1:1000) and surface antibodies (1:200) in PBS at RT for 20 minutes, wash the cells using sterile flow staining buffer. Sort cells using flow cytometric analysis [61], excluding non-viable cells, CD14+ cells, and CD19+ cells; and including cells that are positive for all other markers in the antibody panel.

4.10.2 Reprogramming of T_N and T_{SM} cells into T-iPSCs

Sorted T_N or T_{SM} cells were stimulated before reprogramming for 48 hours by culturing 300,000 cells at a 3:1 ratio with anti-CD3/CD28 beads, 50 IU/mL IL-2, 25 ng/mL IL-7, and 25 ng/mL IL-15 in 6-well plates (2mL/well). Medium: 10% FBS/1% Pen-Strep RPMI 1640. After 48 hours of stimulation, cells were transduced with CytoTune 2.0 Sendai

reprogramming vectors at an MOI of 10. After 24 hours, the media was changed and cells were plated on to mouse embryonic fibroblasts (MEFS) and observed daily for colony formation. Individual colonies were selected and then cultured until loss of episomal Sendai vector was confirmed.

4.10.3 Characterization and validation of pluripotency of T-iPSCs

Genome analyses (nucleic acid isolation, quantitative polymerase chain reactions, bisulfite sequencing, and karyotypes) were performed with standard techniques as described in [215]. For live staining, the TRA-1-60 antibody (1:400, Millipore, Billerica, MA) and secondary antibody Alexa 488-conjugated anti-mouse IgM (1:400, Invitrogen) were diluted in human embryonic stem cell medium and added into the culture plate. The plate was incubated at 37°C for 1 hour before medium was changed to fresh conditioned medium. TRA-1-60+ colonies were identified under a fluorescence microscope. For immunofluorescence evaluations, iPSCs grown on feeder cells in chamber slides were fixed with 4% paraformaldehyde for 15 minutes. If nuclear permeation was needed, cells were treated with 0.2% TritonX (Sigma-Aldrich, St. Louis, MO) in phosphate buffered saline for 30 minutes. Cell preparations were blocked in 3% bovine serum albumin in phosphate buffered saline for 2 hours, and incubated with primary antibody overnight at 4°C. The following antibodies were used: TRA1-60 (clone MAB4360, 1:400), TRA1-81 (clone MAB4381, 1:400), SSEA4 (clone MAB4304, 1:100) and SSEA3 (clone MAB-4303, 1:100) from Millipore; NANOG (clone EB06860, 1:100) from Everest Biotech, Upper Heyford, Oxfordshire, UK; OCT3/4 (clone AB27985, 1:200) from ABCAM, Cambridge, MA; and SOX2 (clone 630802, 1:500) from Biolegend, San Diego, CA. All secondary antibodies used were Alexa Fluor Series from Invitrogen (all 1:500) for 1 hour at room temperature. Images were taken using confocal microscope (Olympus BX61). Direct alkaline phosphatase activity was analyzed using Alkaline Phosphatase Staining Kit used according to the manufacturer's recommendations (Millipore). For teratoma formation, young adult NOG mice were injected with 1 million cells re-suspended in a mixture of DMEM/F12 Matrigel (BD Biosciences), and type IV collagen (ratio 2:1:1, 40 μL per mouse) into the right quadriceps muscle. Tumors were harvested in 3–8 weeks and cryopreserved at -80°C in optimal cutting temperature medium (Sakura Finetek USA).

4.10.4 Preparation, sequencing, and quantification of the accessible genome with ATAC-seq

50,000 cells per biological replicate were sorted. Cells were washed once with 50μL of cold PBS buffer, then lysed with 50 μl of cold lysis buffer (10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$ and 0.1% IGEPAL CA-630). Lysed nuclei were incubated in the transposase reaction mix as previously described [24] at 37°C for 30 minutes. Immediately following incubation, the sample was then purified the DNA using Qiagen MinElute Kit according to

the manufacturer's protocol and eluted transposed DNA in 20 µl Elution Buffer (10mM Tris buffer, pH 8). Purified DNA was stored at -20°C. To amplify transposed DNA fragments and barcoded the libraries, Illumina Nextera Index Kit (FC-121-1011) was used. The following was combined in a PCR tube: 20 µl Transposed DNA, 2.5 µl Nextera PCR Primer with barcode 1, 2.5 µl Nextera PCR Primer with barcode 2, 25 µl NEBNext High-Fidelity 2x PCR Master Mix (New England Labs Cat. #M0541) to 50 µl total volume. PCR cycle as follows: (1) 72°C for 5 min, (2) 98°C 30 sec, (3) 98°C 10 sec, (4) 63°C 30 sec, (5) 72°C, 1 min, (6) Repeat steps 3-5, 12x (7) Hold at 4°C. AMPure XP Beads were used to purify the ATAC-seq libraries with a 1:1 PCR product to Beads ratio. Purified ATAC-seq libraries were eluted in 40 µl Elution Buffer.

ATAC-seq fragments were size-selected for fragments between 115 and 600 bp. ATAC libraries were ligated with Nextera sequencing primers using Polymerase Chain Reaction (PCR), as described [188]. KAPA Library Quantification Kit for Illumina Platforms (KAPA Biosystems, KK4824) was used for sequencing library quantification according to the manufacturer's protocol with ABI ViiA 7 Real-Time PCR System. NextSeq 500 V2 High Output Kit (Illumina, FC-404-2005, 75 cycles) and Illumina NextSeq 550 system were used for library sequencing. ATAC-seq samples were run with paired-end 37bp reads.

To process ATAC-seq, we first used Trimmomatic [16] to trim primers from reads. We aligned reads to the reference genome (hg19) using bowtie version 2.3.2 [112]. We marked duplicate reads in aligned bam files using PICARD. We used FASTQC [51] to verify the quality of aligned reads.

After alignment of reads, reads aligned to the positive and negative strands were shifted +4bp and -5bp, respectively [25]. For each sample, we called peaks using MACS2 (version 2.1.0) [239], setting the FDR to 0.05. Peaks from all samples were merged by taking the union of all overlapping peaks. Overlapping peaks were merged.

4.10.5 Preparation, sequencing, and quantification of the transcribed genes with RNA-seq

50,000-200,000 cells were sorted per biological replicate. Cells were washed once with 500µL of cold PBS buffer. Cell pellets were resuspended in 300µL of RLT lysis buffer and stored at -80°C. Total RNA was extracted with QIAGEN RNeasy Plus Mini Kit (QIAGEN, 74134). mRNA was isolated with NEBNext Poly(A) mRNA Magnetic Isolation Module (NEB, E7490) and the RNA-seq libraries were generated with NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (NEB, E7760) according to manufacturer's instructions.

KAPA Library Quantification Kit for Illumina Platforms (KAPA Biosystems, KK4824) was used for sequencing library quantification according to the manufacturer's protocol with ABI ViiA 7 Real-Time PCR System. NextSeq 500 V2 High Output Kit (Illumina, FC-404-2005, 75 cycles) and Illumina NextSeq 550 system were used for library sequencing with single end 75bp reads.

RNA-seq samples were aligned to the hg38 genome using bowtie2 version 2.3.2. RSEM

was used to quantify gene expression for all samples [117]. PICARD was used to gather metrics using the command CollectRnaSeqMetrics [164].

4.10.6 Reprogramming of fibroblast cells in FB-iPSCs

To isolate dermal FBs for reprogramming, skin biopsies were obtained following consent from healthy donors. Samples were cut into 2 × 2 mm squares, the tissue was then trapped under a sterile coverslip in a 6-well plate, and cultured in human FB medium: DMEM (Invitrogen) supplemented with 10% FBS (HyClone, Logan, UT) and penicillin/streptomycin (Invitrogen). Every two days the media was changed and at 90% confluency, FB cells were passaged using 0.05% trypsin EDTA. FB reprogramming was achieved as previously published using Sendai virus [214, 213, 57]. The four Yamanaka factor Sendai viral supernatants, OCT4, SOX2, KLF4, and c-MYC were produced by transfecting 293T/17 cells with Lipofectamine 2000 (Invitrogen, Carlsbad, CA) with three plasmids: one cargo plasmid (containing the reprogramming gene), a plasmid expressing the VSV-G envelope gene, and a helper plasmid with the retroviral Gag/Pol gene, and then harvested 48-72hours after transfection. Approximately, 100,000 FBs per well of a 6-well plate were plated and infected with a 1:1:1:1 mix of the 4 retroviral supernatants in the presence of 5μgml-1 of protamine sulfate. Five days later, cells were re-seeded onto feeder layers of irradiated CF1 MEFs. The medium was changed after 24 hours to human embryonic stem cell medium, consisting of DMEM/F12 (Invitrogen) supplemented with 10% KnockOut Serum Replacement (Invitrogen), 2mM GlutaMAX (Invitrogen), 50μM 2-mercaptoethanol (Invitrogen), 1 × nonessential amino acids (Invitrogen), 50 Unitsml-1 penicillin, 50mgml-1 streptomycin, and 10ngml-1 basic FB growth factor (R&D Systems, Minneapolis, MN). Cultures were maintained at 37°C, 5% CO2, with daily medium changes. Morphologic appearance was used to pick colonies, 30-60 days after initial infection. FB-iPSC colonies were verified using karyotype, gene expression array, and immunofluorescence and cultured until loss of episomal Sendai vector was confirmed [160].

4.10.7 Gene set enrichment analysis (GSEA) of differentially expressed genes

We used all MSIGDB gene set annotations (version v6.2) [121] to detect increased expression of a set of genes (the foreground), against a background. To detect increased expression of gene sets in a given condition (condition 1), relative to another condition (condition 2), we set the foreground to be the top 2,000 genes identified by DeSeq2 to have significantly increased expression relative to condition 2 (adjusted p-value < 0.05), where genes are ordered by their absolute log2 fold change. We set the background set to be all genes that have base mean expression of greater than 1 between conditions 1 and 2. We used R package clusterProfiler [234] to collect perform GSEA and use a p-value cutoff of 0.05 to determine whether or not a gene set is significant.

4.10.8 Annotation of differentially accessible regions with rGREAT

We use rGREAT [69] to annotate differentially accessible regions in FB-iPSCs and T-iPSCs. We set the foreground to all genomic regions that are significantly differentially accessible in fibroblasts or T cells both before and after reprogramming. We determine significance as all regions with FDR < 0.05. We set the background to all regions with base mean accessibility greater than 1.0, identified using DeSeq2, both before and after reprogramming.

4.10.9 Identification of fibroblast expressed genes and T cell expressed genes

We identified sets of 'fibroblast expressed genes' by selecting the top 2000 genes that had significantly increased expression in fibroblasts, relative to original T cells and ES cells (GEO accession GSE115046) [83]. We additionally compared to ES cells to remove any genes associated with stemness and pluripotency. We first ran a gene set enrichment analysis (GSEA) to identify annotations that had overall increased expression in original fibroblasts, relative to T cells. This annotations identified annotations related to collagen, the extracellular matrix, and epithelial cells. These annotations are of particular interest because fibroblasts synthesize collagen and the extracellular matrix [150], and are expected to be present in the original fibroblasts. We used all genes in this analysis that were annotated in at least one of these categories to define our set of 'fibroblast expressed genes'. We additionally removed any genes annotated in MSigDB to be associated with pluripotency or stemness.

We similarly determined a set of 'T cell expressed genes" by selecting the top 2000 genes that had significantly increased expression in T cells, compared to both fibroblasts and ES cells. We next ran GSEA of these genes, and further selected genes that were annotated for CD8(+) T cell and immune response, as specified in MSigDB. We additionally removed any genes annotated in MSigDB to be associated with pluripotency or stemness.

4.10.10 Maintenance and differentiation of human iPSCs to hematopoietic progenitors

Human iPSC lines were maintained on Matrigel or Geltrex coated plasticware with feeder free, serum free culture conditions in TeSR1 medium (STEMCELL Technologies) at 37 °C incubator with 5% CO2 and 5% O2. For differentiation, hiPSCs were cultured at around 80-90 % confluency, followed by Rock inhibitor (10 μM, TOCRIS) treatment for minimum 30 minutes. For Embryoid body (EB) generation, undifferentiated hiPSCs were washed once with PBS followed by dissociated with Accutase (STEMCELL Technologies) treatment at 37 °C for 5-10 minutes. Large clusters of cells were resuspended in PBS and centrifuged at 270 X g for 5 minutes. After discarding supernatant, cell pellets were gently resuspended in 90% APEL-differentiation medium + 10% TeSR1 medium (STEMCELL Technologies),

supplemented with BMP-4 (10 ng ml-1) and bFGF (5 ng ml-1) and triturated carefully to generate 5-10 cells size aggregates and cultured in non tissue culture plates. Following 42 h, developing EBs were collected by gravity, and resuspended in APEL-differentiation media containing BMP-4 (10 ng ml-1), bFGF (5 ng ml-1), and CHIR99021 (3 μM, Stemgent), SB431542 (6 μM, Selleck Chemicals). After 24 h, EBs were again collected and resuspended in APEL-differentiation media containing VEGF (20 ng ml-1), bFGF (5 ng ml-1), IL-3 (20 ng ml-1), Flt3L (20 ng ml-1) and SCF (100 ng ml-1) and cultured for another 5-6 days with addition of 5 ml fresh media at mid time point. Cultures were maintained in a 5% CO2/5% O2/90% N2 environment. At day 8/9 EBs were harvested, washed once with PBS and dissociated using 1:1 Accutase and 0.25 % trypsin EDTA mixture until no visible clumps were observed. To break clumps into single cell suspension, they were 2-3 times passed through 18 followed by 23 G needle. Hematopoietic CD34+ cells were enriched using Easy-Sep CD34+ isolation kit (STEMCELL Technologies).

For T lineage differentiation, 1 X 105 purified CD34+ cells were plated onto confluent OP9-DLL4 cells for about 3 to 4 weeks and passaged every 4-5 days as described previously [102]. For passaging OP9-DLL4;Tprog cultures, cells were washed with PBS followed by dissociated with accutase treatment at 37^{o}C for 5 minutes. Cells were detached from the surface and triturated using P1000 pipette tips by several times up and down. Single cells suspension were passed through 40μm filter and centrifuged at 400 X g for 5 minutes. Cell pellets were resuspended in 2-3 ml α-MEM complete media with 20% FBS, 100 ng ml-1 SCF (R&D Systems), 10 ng ml-1 Flt3L (Miltenyie Biotec) and 5 ng ml-1 IL-7 (Miltenyi Biotec) for initial 5 days. From Day 6 onward, cells were cultured in 2-3 ml α-MEM complete media with 20% FBS, 10 ng ml-1 SCF, 10 ng ml-1 Flt3L and 5 ng ml-1 IL-7. All recombinant factors are human.

4.10.11 T specification of iPSC-derived CD34(+) Hematopoietic Progenitors

For T lineage differentiation, 1 X 105 purified CD34(+) cells were plated onto confluent OP9-DLL4 cells for about 3 to 4 weeks and passaged every 4-5 days as described previously [102]. For passaging T progenitor cultures, cells were washed with PBS followed by dissociated with Accutase treatment at 37^{o}C for 5 minutes. Cells were detached from the surface and triturated using P1000 pipette tips by several times up and down. Single cell suspensions were passed through 40μm filter and centrifuged at 400 X g for 5 minutes. Cell pellets were resuspended in 2-3 ml α-MEM complete media with 20% FBS, 100 ng ml-1 SCF (R&D Systems), 10 ng ml-1 Flt3L (Miltenyie Biotec) and 5 ng ml-1 IL-7 (Miltenyi Biotec) for initial 5 days. From Day 6 onward, cells were cultured in 2-3 ml α-MEM complete media with 20% FBS, 10 ng ml-1 SCF, 10 ng ml-1 Flt3L and 5 ng ml-1 IL-7. All recombinant factors are human.

Chapter 5

Self-guarding of MORC3 enables virulence factor-triggered immunity

The innate immune system senses pathogens and initiates the defense of hosts [88]. In response to the innate immune system, pathogens employ virulence factors, which are small molecules that aid in the inhibition of immunity [52]. The guard hypothesis [12] postulates that hosts counteract pathogen virulence factors by guarding critical innate immune pathways such that their disruption by virulence factors provokes a secondary immune response [52]. Although this hypothesis is acknowledged in plants [12, 92], the importance of guard immunity in mammals is less clear [52]. This work describes a unique 'self-guarded' immune pathway in human monocytes, in which guarding and guarded function are united in one protein. This pathway is triggered by ICP0, a virulence factor of Herpes Simplex Virus-1 (HSV-1), which results in the induction of type I interferons (IFN). A CRISPR-screen identified the ICP0 target MORC3 as an essential negative regulator of IFN. Mechanistically, ICP0 degrades MORC3, which leads to de-repression of a regulatory element near the IFNB1 locus to drive anti-viral IFN response. These results suggest a model in which the primary anti-viral function of MORC3 is 'self-guarded' by its secondary IFN-repressing function. Thus, a virus that degrades MORC3 to avoid its primary anti-viral function will trigger a secondary anti-viral IFN response.

In this chapter, we gather measurements of the transcriptome and chromatin accessibility to help characterize the role of a transcription factor, MORC3, in the innate immune system's response to pathogens. In particular, we leverage chromatin accessibility to identify a MORC3 regulated element (MRE) that induces transcription of nearby genes in human monocytes. This project is in collaboration with the Vance Lab at University of California, Berkeley. Primary contributions in this work involve interpretation of changes in the transcriptome, and identification of regulatory regions. We include additional experimental procedures performed by collaborators to provide background and motivation.

5.1 Background

Pathogen-associated-molecular patterns (PAMP) are small, molecular motifs that are associated with a given class of pathogens. To protect host cells from viral infection, the innate immune system employs pattern-recognition-receptors (PRR) to detect PAMPs [88]. Detection of PAMPs by PRRs triggers the activation of multiple signaling cascades in the host immune cells, such as the stimulation of interferons (IFNs) or cytokines [165, 2]. Stimulation of type I IFNs triggers the expression of interferon stimulated genes (ISG), which up-regulates the function of immune cells to resolve pathogens [148].

To avoid these immune defenses, pathogens produce virulence factors that disrupt critical immune pathways [52]. To combat disruption of immune pathways by virulence factors, immunity in plants relies on a secondary line of pathogen detection, which monitors, or guards, the integrity of host immune defense pathways. This guarding of host defense pathways is referred to as the guard hypothesis [12], which results in virulence factor-mediated disruption of the guarded pathway, which triggers activation of a secondary immune response [52]. Plants employ numerous guard proteins that are critical for immunity [92, 227]. Other sensors also guard critical pathways against bacterial virulence factors in mammals [52, 18, 99, 231]. However, the extent to which guard-immunity in mammals extends beyond these sensors remains unclear [52].

Herpes Simplex Virus-1 (HSV-1) is a dsDNA-virus. Its DNA genome is recognized as a PAMP by the cyclic-GMP-AMP synthase (cGAS)–stimulator of interferon genes (STING) PRR-pathway [127, 78]. Downstream of DNA recognition, STING activates TANK Binding Kinase 1 (TBK1) [85] and I-kappa-B kinase ϵ (IKKϵ) [59] that phosphorylate and activate interferon regulatory factors 3 and 7 (IRF3/7), leading to transcriptional induction of type I interferon (IFN) cytokines [78]. IFNs bind to the IFN alpha and beta receptor (IFNAR) to induce transcription of anti-viral interferon-stimulated genes (ISG). In this work, we evaluate guard-immunity through the innate immune response to HSV-1 in human BLaER1 monocytes.

5.2 CRISPR screen identifies MORC3 as a negative regulator of IFN in human monocytes

The primary function of virulence factor ICP0 is to degrade or inactivate host proteins [17]. Therefore, it was hypothesized that ICP0 targets a negative regulator of IFN for degradation (Figure 5.1a), inadvertently triggering an IFN response. To identify negative regulators of IFN in BLaER1 monocytes, a genome-wide CRISPR screen was conducted to identify a substrate of ICP0 that negatively regulates IFN. The screen relied on antibody staining of Virus inhibitory protein, endoplasmic reticulum-associated, interferon-inducible (Viperin). Viperin is an ISG that can be used to report IFN status at single cell resolution (Figure 5.1b).

Cells with spontaneous induction of Viperin were sorted from the bulk population to enrich for sgRNAs whose targets are negative regulators of IFN. sgRNAs targeting the protein

MORC3 induced Viperin expression. We decided to focus on MORC3 as it was previously described to be degraded by ICP0 in fibroblasts [195]. We confirmed that ICP0 was required for MORC3 degradation during HSV-1 infection in BLaER1 monocytes (Figure 5.1d). Independently generated MORC3–/– BLaER1 monocytes (Figure 5.1e) displayed spontaneous IFN induction, measured by IFNB1.

We next collected RNA-seq from MORC3–/– monocytes and WT stimulated monocytes (mCherry) to understand how the degradation of MORC3 changes the global transcriptome. We also collected samples for virulence factor induced monocytes (two virulence factors: ICP0 and E4ORF3) to understand how the degradation of MORC3 changes the global transcriptome. Global transcriptomic changes in MORC3–/– monocytes mirrored changes induced by virulence factors, leading to co-clustering of these conditions on the first principal component, generated from gene counts (Figure 5.1f). Furthermore, investigation of expression of ISGs showed similar expression patterns of ISGs in MORC3–/– monocytes and virulence factor induced monocytes (Figure 5.1g). These results identify MORC3 as a negative regulator of IFN in human monocytes and demonstrate that genetic loss or ICP0-mediated degradation of MORC3 leads to a potent IFN response that is independent of PRR-signaling hubs.

5.3 MORC3 represses viral replication and IFN transcription

To investigate how MORC3 represses IFN we next performed RNA-seq of MORC3–/– and MORC3–/– IFNAR1–/– IFNAR2–/– monocytes. We considered MORC3–/– IFNAR1–/– IFNAR2–/– monocytes in particular because IFNs signal through the IFNAR receptor, triggering stimulation of ISGs. We found the majority of transcriptional changes incurred by MORC3 deficiency were due to IFNAR signaling. ISGs that were induced in MORC3–/– monocytes were not upregulated upon IFNAR co-deletion (Figure 5.2d) and MORC3–/– IFNAR1–/– IFNAR2–/– cells clustered together with IFNAR1–/–IFNAR2–/– cells (Figure 5.2e). This indicates that MORC3 is not required to directly repress antiviral ISGs. However, a small number of genes, including IFNB1, were still de-repressed in MORC3–/– IFNAR1–/– IFNAR2–/– monocytes, which suggested they may be direct targets of MORC3 repression (Figure 5.2d). The most significantly de-repressed genes in MORC3–/– IFNAR1–/– IFNAR2–/– monocytes were MLLT3, IFNB1 and FOCAD, which, remarkably, are clustered together within a short section of chromosome 9 (chr9:20,329,000-21,086,000; Figure 5.2i). Almost all protein-coding genes within this region were de-repressed in MORC3–/– cells, while adjacent genes in both directions were not regulated (Figure 5.2f). Given that these genes have different promoters and are not normally co-regulated or functionally related, we hypothesized that MORC3 acts in a locus-specific manner to regulate gene expression. Consistent with this, a retroviral-based randomly integrated IFNB1-promoter-luciferase reporter was only activated by DNA-STING activation but was not activated in the absence

of MORC3 (Figure 5.2g). As expected, the endogenous IFNB1 gene was activated both by STING signaling and MORC3 deficiency (Figure 5.2h). Together, these results indicate that MORC3 acts to repress the IFNB1 locus rather than the IFNB1 promoter, whereas canonical PRR-mediated induction of IFNB1 is promoter-dependent and locus-independent.

We next asked how the location-dependent repression mediated by MORC3 regulates expression of anti-viral ISGs. There are 17 different type I IFN genes (IFNB1, IFNE, IFNK, IFNW1, and thirteen IFNA genes) that cluster together on chromosome 9 but encode proteins that vary in cell-type expression, kinetics of induction, and receptor affinity. The most studied type I IFN gene is IFNB1 because it is dominantly induced in IFN-producing cells. However, ISG induction downstream of STING activation in BLaER1 monocytes did not rely solely on IFNB1, as revealed by IFNAR-dependent induction of ISGs such as RSAD2, CXCL10 in an IFNB1-independent manner in response to foreign DNA (Figure 5.3a). Thus, BLaER1 monocytes are competent to make IFNs other than IFNB1. In contrast, IFNB1 is the only IFN-gene within the MORC3-repressed genomic region on chromosome 9 (Figure 5.2d) and is thus the only IFN-gene that is de-repressed in MORC3–/– cells (Figure 5.3b). Consequently, monocytes required IFNB1 to activate ISGs in the absence of MORC3 (Figure 5.3c, d). MORC3–/–IFNB1–/– cells clustered together with MORC3–/–IFNAR1–/–IFNAR2–/– on the PCA plot of RNA-seq data (Figure 5.3e). As expected, virulence factor-mediated activation of the MORC3 pathway also resulted in the locus-specific induction of IFNB1 and its neighboring genes (Figure 5.3f). We conclude that MORC3 acts narrowly and selectively on the IFNB1 locus and does not regulate other type I IFN genes. Thus, in contrast to canonical PRR-signaling, activation of the MORC3 pathway by viral virulence factors uniquely relies on IFNB1 for anti-viral ISG induction (Figure 5.3g).

5.4 Identification of regulatory regions of IFNB1 in BLaER1 monocytes

Because MORC3 is known to interact with H3K4me3 [118] and repress IFN, we next sought to understand how MORC3 was repressing IFNB1. Specifically, we sought to verify if repression of IFNB1 was related to changes in chromatin accessibility in the genomic region containing IFNB1, and whether a change in accessibility was driving de-repression of IFNB1 in MORC3–/–IFNAR1–/– IFNAR2–/– monocytes. To investigate changes in chromatin accessibility in the absence of MORC3, we collected ATAC-seq from IFNAR1–/–IFNAR2–/– and MORC3–/–IFNAR1–/–IFNAR2–/– monocytes. When evaluating genome-wide changes in accessibility between conditions, most strongly differentially accessible regions with significant de-repression in MORC3–/–IFNAR1–/– IFNAR2–/– monocytes were contained near IFNB1(Figure 5.4(a)). However, regions outside this region, including promoters of IGF2BP3 and ZNF239, were also significantly differential between conditions. To identify regions that may be regulatory for de-repression of IFNB1, we filtered out promoter regions from our

analysis (Figure 5.4(b)). After this filter, there were two outlier peaks with significant enrichment in MORC3–/–IFNAR1–/– IFNAR2–/– monocytes located near IFNB1, contained in the intron of the FOCAD gene (FDR of 1.54e-24 and 8.46e-20). We refer to this region a s MORC3 regulated element, as its accessibility is induced in the absence of MORC3. Figure 5.4(c) shows ATAC-seq coverage at the MRE in the FOCAD intron, showing significant de-repression in MORC3–/–IFNAR1–/– IFNAR2–/– monocytes.

To confirm these results, we collected RNA-seq that knocked out the MRE to determine whether IFNB1 expression was dependent on this potential enhancer in the absence of MORC3. Indeed, IFNB1 expression was significantly reduced in MRE-/- MORC3-/- monocytes (Figure 5.4(d)). These results suggest that we have identified a region that is regulatory for expression of IFNB1, and is a component of the MORC3 pathway.

We next leveraged Epitome [144], discussed in Section 2.3.1, to predict binding sites of previously characterized TFs to better understand regulation of this identified peak. We predicted binding for 97 TFs near IFNB1 in IFNAR1–/–IFNAR2–/– and MORC3–/–IFNAR1–/– IFNAR2–/– monocytes, leveraging ATAC-seq from both conditions as input to Epitome as a measurement of cell type specificity. We used a t-test to identify whether predictions of binding were significantly different between IFNAR1–/–IFNAR2–/– and MORC3–/–IFNAR1–/– IFNAR2–/– conditions, and corrected resulting p-values for FDR. Epitome identified enrichment of various TFs in MORC3–/–IFNAR1–/– IFNAR2–/– monocytes in regions near IFNB1 (chr9:20,500,000-21,500,000), including IKZF1, TEAD4, and TAF1. These results were not verified with ChIP-seq or CUT&RUN.

5.5 Characterization of MORC3 binding patterns in BLaER1 monocytes

Although we have identified that MORC3 is a negative regulator of IFN in monocytes, the mechanism to which MORC3 regulates IFN is unknown. In particular, we have identified a regulatory region in the FOCAD intron (Figure 5.4(b)) that is de-repressed in the absence of MORC3. We therefore sought to determine whether MORC3 was directly repressing this region. We collected CUT&RUN for MORC3 in IFNAR1–/– IFNAR2–/– monocytes to identify MORC3 binding sites. We used processing pipelines described in Section 3.2 to process CUT&RUN samples. We generated a knock-in cell line that expresses 3xFLAG tagged MORC3 at endogenous levels, and used FLAG M2 antibody, as previously described [67]. However, these experiments failed, with the identification of insufficient numbers of peaks in IFNAR1–/– IFNAR2–/– monocytes, and poor enrichment within identified peaks (percent reads in peaks ranging from 0.8%-0.7%).

We therefore ran ChIP-seq to identify binding sites of MORC3, using MORC3 antibody in IFNAR1–/– IFNAR2–/– monocytes. Although it has been previously suggested that MORC3 co-localizes with H3K4me3 through identification of MORC3 binding sites using the MORC3 antibody [118], we find that the MORC3 antibody may be interacting with

H3K4me3 (Figure 5.6(a)). This hypothesis would make it difficult to identify true MORC3 peaks when using the MORC3 antibody.

5.6 Endogenous retroviruses are not driving de-repression of IFN in MORC3–/– IFNAR1–/– IFNAR2–/– monocytes

Although we have shown that MORC3 represses IFN, previous literature has suggested that MORC3 may be inhibiting endogenous retroviruses (ERVs) [67]. ERVs are discussed in detail in Section 3.1. In a recent study, Groh et.al. showed that MORC3 binds ERV sequences in mouse embryonic stem (ES) cells, silencing ERVs [67]. Thus, in the absence of MORC3, increased expression of ERVs may be directly regulating IFN response. We therefore used protocol discussed in Section 3.1 to quantify expression of ERVs from RNA-seq in four conditions: IFNAR1–/–IFNAR2–/– mCherry, MORC3–/–, MORC3–/–IFNAR1–/–IFNAR2–/–, and MORC3–/–IFNB1–/– monocytes. We next identified ERVs significantly enriched in the former three conditions, relative to IFNAR1–/–IFNAR2–/– mCherry. Across these conditions, three ERVs had shared enrichment in all three conditions, relative to mCherry (Figure 5.5(a)). These ERVs included MSR1, HERVL74-int, and LTR12B.

We additionally extended our analysis to quantify ERVs using RepEnrich2 [37], designed to quantify genome-wide levels of repetitive elements. We chose RepEnrich2 in particular because it was used by Groh et al. [67] to quantify ERVs in the absence of MORC3 in ES cells. Figures 5.5(b) shows that no enrichment of ERVs between IFNAR1–/–IFNAR2–/– mCherry and MORC3–/–IFNAR1–/–IFNAR2–/– monocytes was identified. However, DeSeq2 identified multiple ERVs that were significantly enriched (FDR < 0.05) in both conditions, suggesting that there was no preference of ERV enrichment in MORC3–/–IFNAR1–/–IFNAR2–/– monocytes.

5.7 Discussion

These results identify a novel innate immune sensing mechanism that detects the enzymatic activity of virulence factors from DNA viruses. This pathway does not utilize a PAMP-sensing receptor akin to a PRR; nor does it depend on canonical PRR signaling components that many viruses have evolved to disrupt or inhibit. Instead, it employs the self-guarded protein MORC3, whose bi-functionality allows detection of pathogen-encoded enzymatic activities. The primary function of MORC3 appears to be to inhibit replication of HSV-1 [138, 204]. To escape restriction by MORC3, DNA viruses employ virulence factors, which may degrade MORC3. In response, we propose MORC3 evolved a secondary function: locus-specific repression of the IFNB1 gene. This secondary function allows activation of anti-viral IFN upon virulence factor-mediated perturbation of ND10 nuclear bodies and MORC3.

Employing a single protein to repress both viral gene expression and the IFNB1 locus may provide a significant barrier against viruses that seek to selectively escape the repression of viral genes without also triggering IFNB1 expression. It is conceivable that MORC3 exerts its two different functions by executing one unifying molecular activity, namely, transcriptional repression. Various repressive activities have been proposed for the MORC gene family, including DNA methylation [125], H3K9-methylation [207], H3.3 incorporation [67] and DNA compaction [103]. In the future, it will be of interest to determine how MORC3 represses both viral and IFNB1 gene expression and how repressive activity is targeted to specific loci.

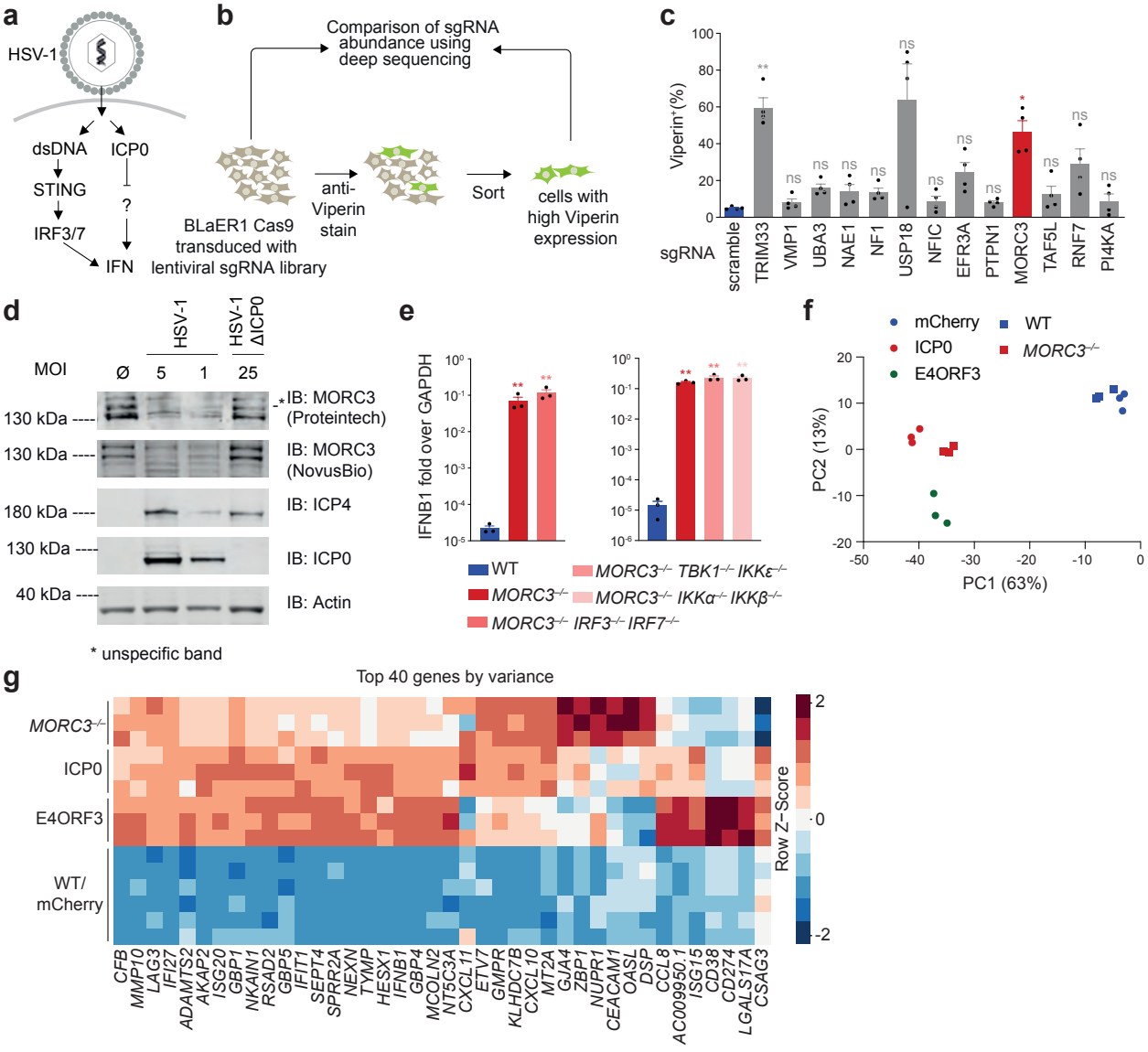


Figure 5.1: MORC3 is a novel negative regulator of IFN. a) Hypothesized mechanism in which ICP0 degrades a negative regulator of IFN. b) Schematic of genome wide CRISPR screen to identify negative regulator of IFN. c) Cas9-expressing BLaER1 cells were transduced with individual sgRNAs targeting and Viperin expression was analyzed by FACS. Mean+SEM of n=4. * $p < 0.05$; ** $p < 0.01$; ns = not significantly different than scramble sgRNA, tested by one-way ANOVA and Dunnett's post hoc test. d) IFNAR1–/–IFNAR2–/–STING–/–SP100–/– BLaER1 monocytes (lacking factors that would otherwise restrict △ICP0 mutant virus) were infected with HSV-1 for 3h. One representative immunoblot of two is shown. e) Gene expression of BLaER1 monocytes is shown as mean ± SEM of n=2-3 from one representative clone or two (multiple KOs) or one clone (WT and MORC3-/-). ** $p < 0.01$; ns = not significantly different than WT, tested by two-way ANOVA and Dunnett's post hoc test. f, g) Transcriptional changes in BLaER1 monocytes as detected by RNA-seq in three independent experiments are depicted by PCA analysis (f) and heatmap (g).

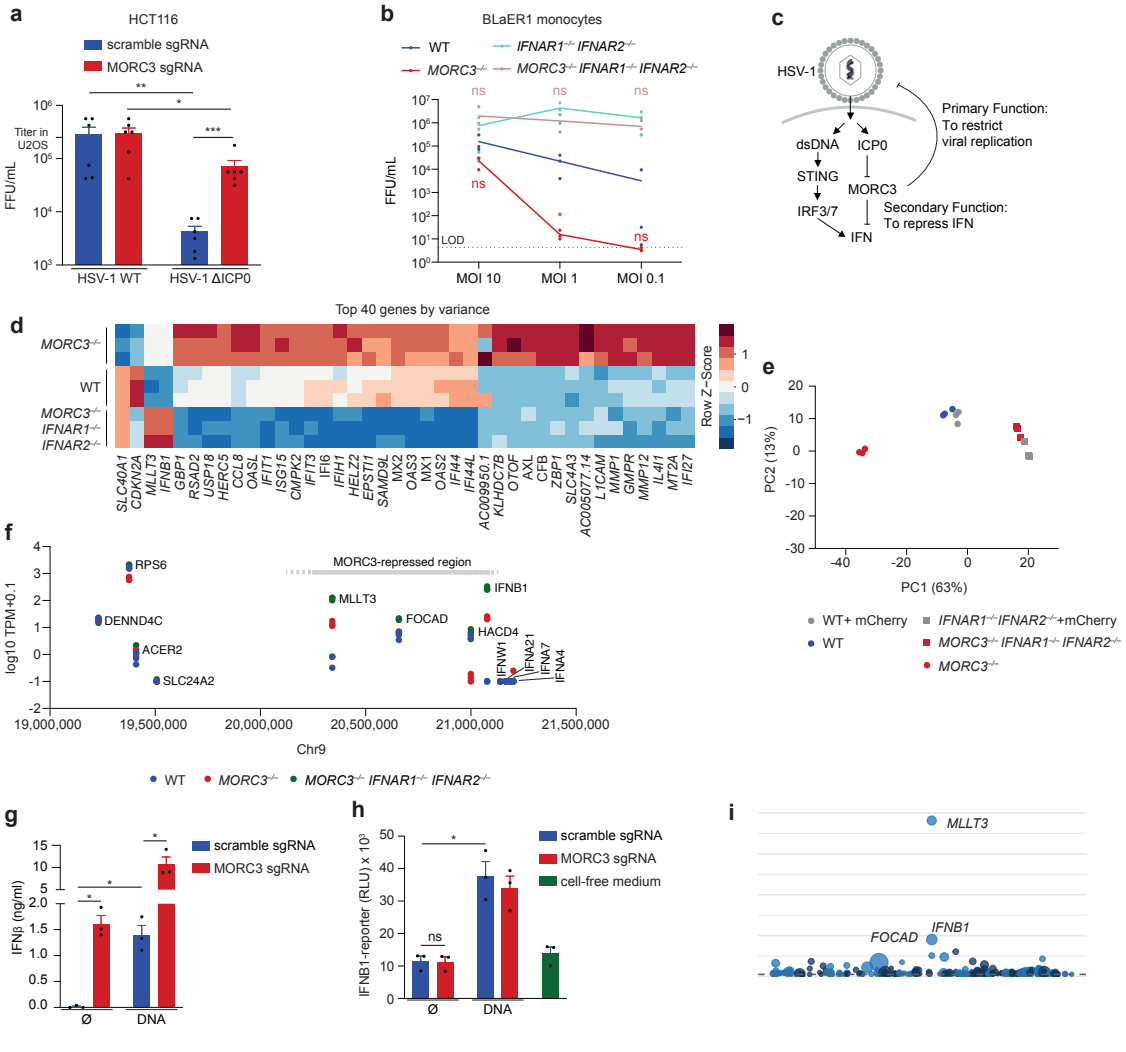

a
HCT116
scramble sgRNA
MORC3 sgRNA
Titer in U2OS
FFU/mL
HSV-1 WT
HSV-1 ΔICP0
b
BLaER1 monocytes
WT
IFNAR1-/- IFNAR2-/-
MORC3-/-
MORC3-/- IFNAR1-/- IFNAR2-/-
FFU/mL
LOD
MOI 10
MOI 1
MOI 0.1
c
HSV-1
dsDNA
ICP0
STING
MORC3
IRF3/7
IFN
Primary Function: To restrict viral replication
Secondary Function: To repress IFN
d
Top 40 genes by variance
MORC3-/-
WT
MORC3-/- IFNAR1-/- IFNAR2-/-
Row Z-Score
SLC40A1 CDKN2A MLLT3 IFNB1 GBP1 RSAD2 USP18 HERC5 CCL8 OASL IFIT1 ISG15 CMPK2 IFIT3 IFI6 IFIH1 HELZ2 EPSTI1 SAMD9L MX2 OAS3 MX1 OAS2 IFI44 IFI44L AC009950.1 KLHDC7B OTOF AXL CFB ZBP1 SLC4A3 AC005077.14 L1CAM MMP1 GMPR MMP12 IL4I1 MT2A IFI27
e
PC2 (13%)
PC1 (63%)
WT+ mCherry
WT
IFNAR1-/- IFNAR2-/- +mCherry
MORC3-/- IFNAR1-/- IFNAR2-/-
MORC3-/-
f
log10 TPM+0.1
MORC3-repressed region
RPS6
DENND4C
ACER2
SLC24A2
MLLT3
FOCAD
IFNB1
HACD4
IFNW1
IFNA21
IFNA7
IFNA4
Chr9
WT
MORC3-/-
MORC3-/- IFNAR1-/- IFNAR2-/-
g
IFNβ (ng/ml)
scramble sgRNA
MORC3 sgRNA
Ø
DNA
h
IFNB1-reporter (RLU) x 10³
scramble sgRNA
MORC3 sgRNA
cell-free medium
ns
Ø
DNA
i
MLLT3
FOCAD
IFNB1

Figure 5.2

Figure 5.2: a) The titer of HSV-1 stocks at 2.5×105 U2OS-FFU/ml was determined on HCT116-Cas9 cells that were transduced with the indicated sgRNAs. Mean ± SEM of n=6. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; ns = not significantly different, tested by paired, two-sided t-test. b) BLaER1 monocytes were infected with HSV-1 and viral progeny were quantified after 48h. Mean (line) and individual values of three independent experiments. LOD = limit of detection. * $p < 0.05$; ns = not significantly different than the corresponding MORC3 sufficient condition, tested by one-way ANOVA and Dunnett's post hoc test. c) Two proposed functions of MORC3 that allow self-guarding. d, e) Transcriptomic changes in BLaER1 monocytes as detected by RNA-seq in three independent experiments: heatmap of log normalized counts of the top 40 most variable genes (column normalized) (a) or PCA of variance stabilizing transformed counts (e). f) Log transcripts per million (TPM) of genes in BLaER1 monocytes of genes clustered near IFNB1 on chromosome 9 as detected by RNA-seq. Depicted are means of three independent experiments. All protein coding genes within this region are depicted. g, h) STAT1–/–STAT2–/–BLaER1-Cas9 expressing a randomly integrated IFNB1-promoter-Luciferase reporter were transduced with the indicated sgRNAs and stimulated with cytosolic DNA for 24h. Luciferase signal and IFNβ secretion is depicted as mean ± SEM of n=3. $p < 0.05$; ns = not significantly different, tested by paired, two-sided t-test. i) Adjusted p-values of differentially expressed genes between MORC3–/– IFNAR1–/– IFNAR2–/– and IFNAR1–/– IFNAR2–/– mCherry expressing cells are depicted. Data point sizes represent normalized effect size, calculated as the effect size multiplied by the normalized mean counts of all MORC3–/– IFNAR1–/– IFNAR2–/– and IFNAR1–/– IFNAR2–/– expressing cells.

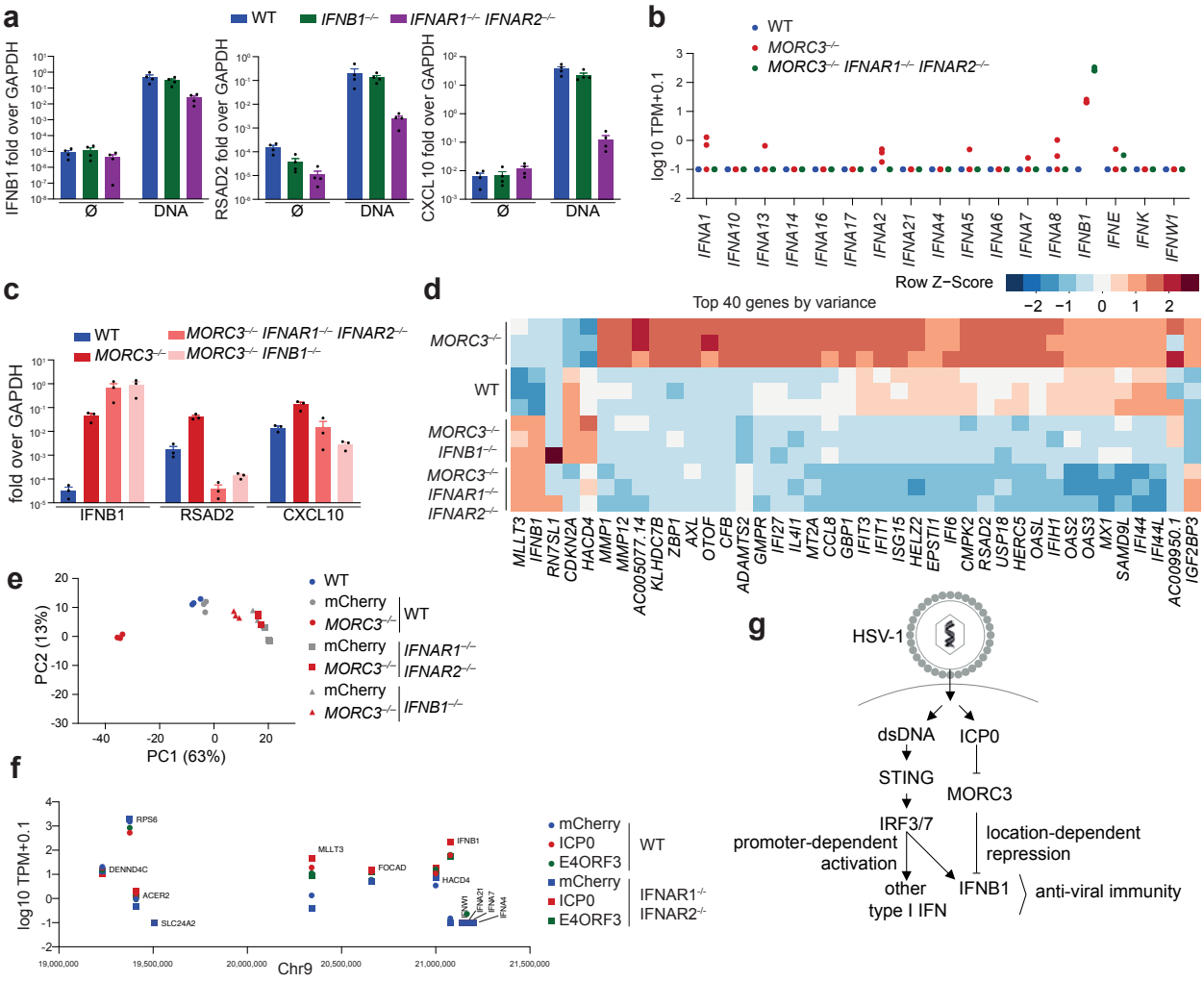


Figure 5.3: Positional repression of IFNB1 explains IFN de-repression in MORC3 deficient cells a) BLaER1 monocytes of the indicated genotypes were transfected with DNA for 12h. Gene-expression as measured by q-RT-PCR of one representative clone of two per genotype is depicted as mean ± SEM of four independent experiments. b) Log transcripts per million (TPM) of IFN genes in BLaER1 monocytes were detected by RNA-seq in three independent experiments. c) Gene expression as measured by q-RT-PCR of BLaER1 monocytes of the indicated genotypes is shown as mean ± SEM of n=3 from one representative clone of two (multiple KOs) or one clone (WT and MORC3−/−). d-e) Transcriptomic changes in BLaER1 monocytes as detected by RNA-seq in three independent experiments are depicted by heatmap analysis (d) and PCA (e). f) Mean log transcripts per million (TPM) of a gene cluster at chromosome 9 in BLaER1 monocytes . Mean TPM is calculated for each condition across three experiments. All protein coding genes within this region are depicted. g) Overview of virulence-factor-triggered immune signaling.

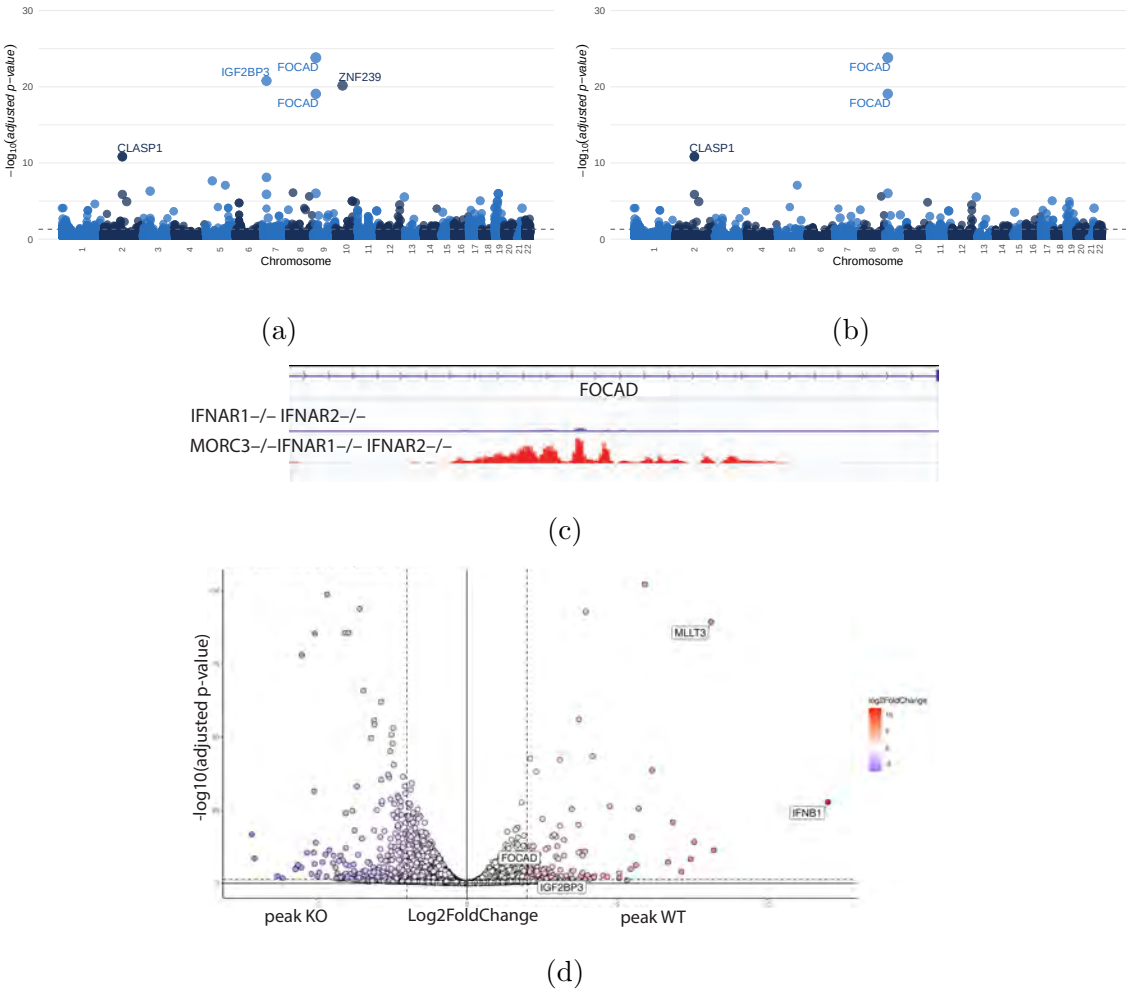


Figure 5.4: Identification of regulatory region for IFNB1 expression. (a) Genome-wide adjusted p-values (FDR) for all differentially accessible regions identified from ATAC-seq between IFNAR1–/– IFNAR2–/– and MORC3–/–IFNAR1–/– IFNAR2–/– monocytes. All regions displayed indicate enrichment in MORC3–/–IFNAR1–/– IFNAR2–/– monocytes. (b) Genome-wide adjusted p-values (FDR) differentially accessible regions not overlapping promoter regions, identified from ATAC-seq between IFNAR1–/– IFNAR2–/– and MORC3–/– IFNAR1–/– IFNAR2–/– monocytes. All regions displayed indicate enrichment in MORC3–/–IFNAR1–/– IFNAR2–/– monocytes. (c) Pileup of ATAC-seq reads for IFNAR1–/– IFNAR2–/– and MORC3–/– IFNAR1–/– IFNAR2–/– monocytes at the FOCAD intron (chr9: 20,972,654 - 209,751,43). (d) Log2 fold change and adjusted p-values (FDR) for all differentially expressed genes between FOCAD intron-/- MORC3-/- STAT1-/- STAT2-/- (peak KO) and MORC3-/- STAT1-/- STAT2-/- (peak WT) monocytes. Red indicates enrichment in peak WT, and blue indicates enrichment in peak KO.

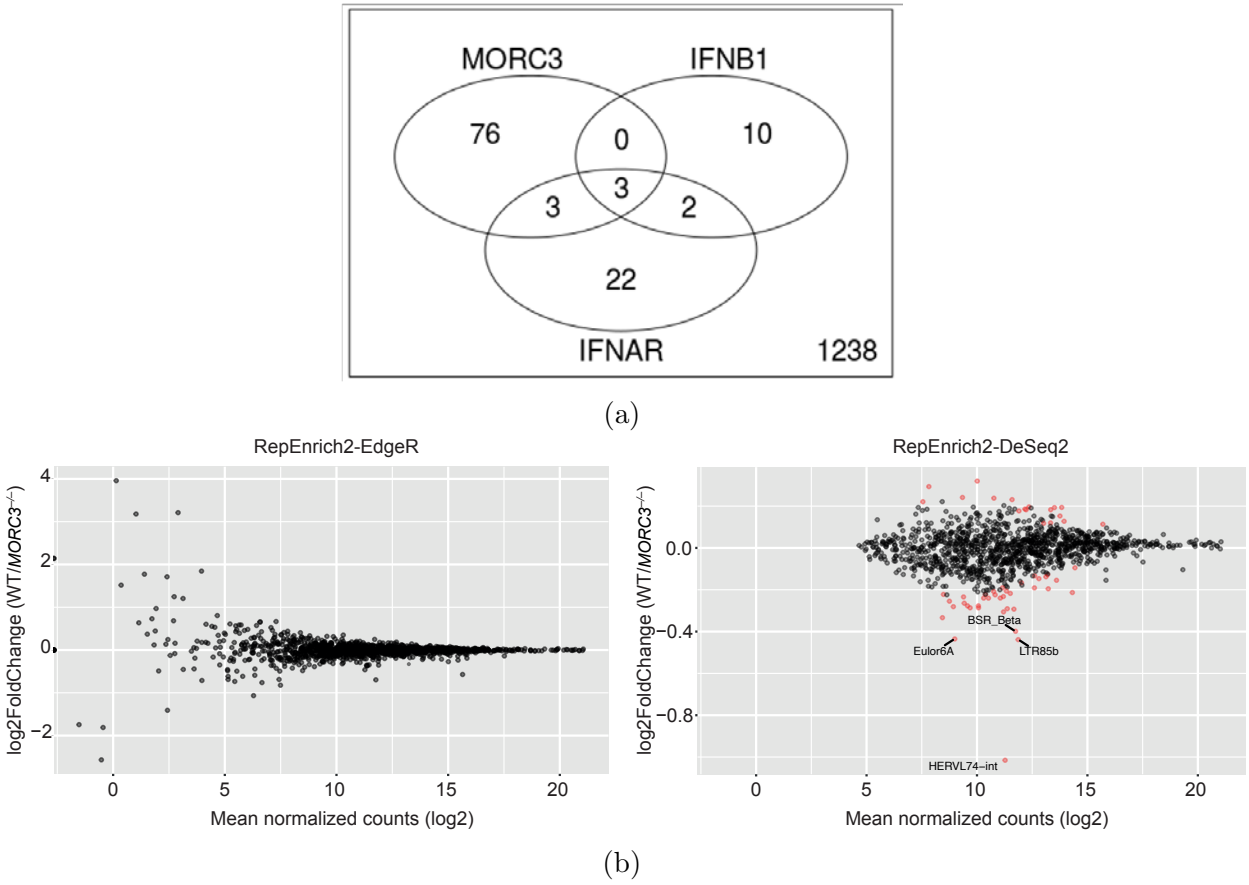
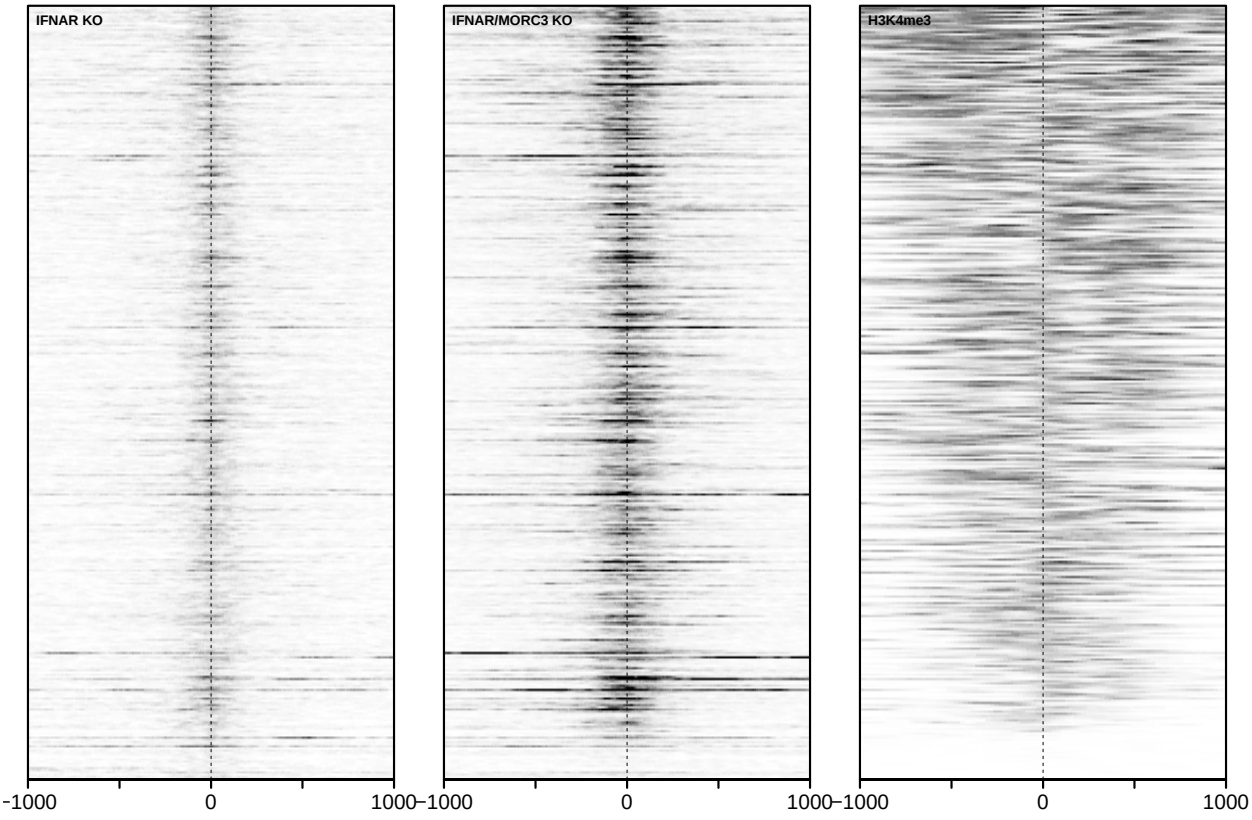


Figure 5.5: Quantification of expression of ERV families in IFNAR1–/–IFNAR2–/– mCherry and MORC3-/- monocytes. (a) Overlap of shared enrichment of ERVs in MORC3–/– IFNAR1–/– IFNAR2–/–, MORC3–/– IFNB1–/–, and MORC3-/- monocytes when compared to IFNAR1–/– IFNAR2–/– mCherry samples. (b) ERV family expression was quantified with RepEnrich2 in IFNAR1–/– IFNAR2–/– mCherry (WT) vs MORC3–/– IFNAR1–/– IFNAR2–/– monocytes. ERVs with an FDR <0.05 are highlighted in red. (left) Using EdgeR [173] did not detect any differential regulation of ERVs upon MORC3 deficiency. (right) DeSeq2 [126] analysis suggested minimal up- and down-regulation of ERV families.



(a)

Figure 5.6: (a) Heatmaps of normalized read counts for 357 peaks identified in IFNAR1-/- IFNAR2-/- monocytes. Each row is centered around a peak summit, displaying counts 2000bp around the peak summit. Reads were normalized for library size. Color scale of IFNAR/MORC3 KO (IFNAR1-/- IFNAR2-/- MORC3-/-) is relative to to IFNAR KO (IFNAR1-/- IFNAR2-/-). H3K4me3 is taken CD14++ CD16- monocytes from blood (accession GSM1320313).

5.8 Methods and Materials

All experimental methods discussed in this Section were performed by the Vance Laboratory at University of California, Berkeley.

5.8.1 Cell culture

BLaER1, U937 and THP-1 cells were cultured in RPMI Medium 1640 supplemented with L-glutamine, sodium pyruvate, 100U/ml penicillin-streptomycin (Thermo Fisher) and 10% (v/v) FCS (Omega Scientific). HEK293T, U2OS and HCT116 cells were cultivated in DMEM Medium (Thermo Fisher) containing the same supplements. 1.4 million BLaER1 cells per well of a 6-well plate were trans-differentiated into monocytes for 5-6 days in medium containing 10 ng/ml of hrIL-3, 10 ng/ml hr-CSF-1 (M-CSF) (both PeproTech) and 100 nM β-Estradiol (Sigma-Aldrich) as previously described [59, 58]. 1.4 million THP-1 and U937 cells per well of a 6-well plate were differentiated overnight with 100 ng/ml PMA (Sigma-Aldrich). STING-deficient and corresponding control THP1 and U937 cells were a gift from Dan Stetson (University of Washington). BLaER1 cells were a gift from Thomas Graf (CRG, Barcelona, Spain) and Veit Hornung (LMU Munich, Germany). U2OS cells were a gift from Robert Tjian and Xavier Darzacq (UC Berkeley). THP1 cells were from ATCC. U937 cells were from the UC Berkeley Cell Culture Facility. HCT116 cells were a gift from David Raulet (UC Berkeley).

5.8.2 Cell stimulation

For activation of the cGAS-STING pathway, 3.2µg of UltraPure™ Salmon Sperm DNA (Thermo Fisher) or 3.2µg of 2'3' cGAMP (Invivogen) was complexed with 8µl Lipofectamine 2000 (Thermo Fisher) according to the manufacturer's protocol in Opti-MEM Reduced Serum Media (Thermo Fisher) and added to 1.4 million cells per well of a 6-well plate for 3h or the indicated time. PRRs were activated with 200 ng/ml LPS-EB ultrapure from E. coli O111:B4 (Invivogen) or 500ng/ml R848 (Invivogen). For activation of doxycycline-inducible trans-gene expression, cells were stimulated with 1µg/ml doxycycline hyclate (Sigma-Aldrich) for 24h.

5.8.3 HSV-1 infection

BACs of △ICP0 HSV-1 and corresponding WT strain were a gift from Bernard Roizman (University of Chicago). BAC DNA was prepared from a mono-clonal transformant and transfected into U2OS cells using Lipofectamine 2000 (Thermo Fisher). Virus was propagated, harvested and frozen as described [15]. Viral progeny were titered from cell-free supernatants by TCID50 using 8 replicates per dilution. U2OS cells were used for titering if not otherwise indicated, and FFU/ml was calculated by the Spearman & Kärber algo-

rithm. Myeloid cells were infected by adsorbing virus of appropriate MOI in FCS free RPMI Medium 1640 for 1h. Subsequently, medium was changed to complete RPMI Medium 1640.

5.8.4 Quantification of gene expression

Gene expression was quantified by RT-qPCR. RNA was isolated with E.Z.N.A. Total RNA kit I (Omegabiotek) and 0.5-1µg RNA was treated with RQ1 RNase-free DNase (Promega) in presence of RNasin plus Ribonuclease Inhibitor (Promega). RNA was reverse transcribed with Superscript III reverse transcriptase (Invitrogen). SYBRGreen dye (Thermo Fisher Scientific) was used for quantitative PCR assays and analyzed with a real-time PCR system (StepOnePlus; Applied Biosystems). All gene expression values were normalized to GAPDH and are depicted as $2^{-}\triangle$Ct (Ct_{target} – Ct_{GAPDH}).

RSAD2.fwd CAACTACAAATGCGGCTTCT
RSAD2.rev ATCTTCTCCATACCAGCTTCC
CXCL10.fwd TCTGAATCCAGAATCGAAGG
CXCL10.rev CTCTGTGTGGTCCATCCTTG
GAPDH.fwd GAGTCAACGGATTTGGTCGT
GAPDH.rev GACAAGCTTCCCGTTCTCAG
IFNB1.fwd CAGCATCTGCTGGTTGAAGA
IFNB1.rev CATTACCTGAAGGCCAAGGA
MLLT3.fwd GAGCACAGTAACATACAGCA
MLLT3.rev GGCAAAATGAAACCAGCATA

5.8.5 Immunoblotting

Whole cell lysates were prepared by lysing cells in 50mM Tris pH7.4, 50mM NaCl, 2mM MgCl2, 0.5% NP40, 25U/ml Benzonase® Nuclease (Millipore Sigma) and Complete Mini EDTA-free Protease Inhibitor (Roche) for 20 min on ice. Laemmli buffer was added to a final concentration of 1× and lysates were boiled at 95°C for 10 minutes. Proteins were separated with denaturing PAGE and transferred to Immobilon-FL PVDF membranes (Millipore Sigma). Membranes were blocked with Li-Cor Odyssey blocking buffer. Primary antibodies were added and immunoblots incubated overnight. Primary antibodies used were anti-β-Actin (C4) (Santa Cruz, sc-47778), anti-HSV-1 ICP4 (H943) (Santa Cruz, sc-69809), anti-HSV-1 ICP0 (11060) (Santa Cruz, sc-53070), anti-TBK1 (D1B4) (Cell Signaling, #3504), anti-IKKϵ (Cell Signaling, #2690), anti-IRF-3 (D83B9) (Cell Signaling, #4302), anti-IRF-7 (Cell Signaling, #4920), anti-MORC3 (NovusBio, NBP1-83036), anti-MORC3 (Proteintech, 24994-1-AP). Appropriate secondary IRDye®-conjugated antibodies (Li-Cor) were used and immunoblots were imaged using the Li-Cor Odyssey platform.

5.8.6 CRISPR/Cas9 mediated gene targeting

Monoclonal gene deficient BLaER1 cells were generated as follows. Briefly, sgRNAs specific for the indicated genes, were designed to target an early coding exon of the respective gene with minimal off-targets and high on-target activity using ChopChop [110]. U6-sgRNA-CMV-mCherry-T2A-Cas9 plasmids were generated by ligation-independent-cloning as previously described [182] and BlaER1 cells were electroporated using a Biorad GenePulser device. Automated cell sorting was used to collect mCherry positive cells that were cloned by limiting dilution. Monoclonal cell lines were identified, rearranged and duplicated for genotyping using deep sequencing as previously described [183]. Knockout cell clones contained all-allelic frame shift mutations without any wild type reads. The MRE was deleted using indicated sgRNAs below to induce a 3kb deletion. Two independent knockout single-cell clones were analyzed per genotype, and one representative clone per genotype is shown. For polyclonal gene targeting, cell lines were transduced with lentiCas9-Blast [178], a gift from Feng Zhang (Addgene plasmid # 52962; http://n2t.net/addgene:52962; RRID: Addgene_52962). sgRNAs were designed as above and cloned into lentiGuide-Puro [178], a gift from Feng Zhang (Addgene plasmid # 52963; http://n2t.net/addgene:52963; RRID: Addgene_52963), using ligation-independent-cloning. Cas9-expressing cells were transduced with indicated sgRNA-encoding lenti-viruses.

sgRNA target sites (PAM is highlighted in bold):
STING GCGGGCCGACCGCATTTGGG**AGG**
TBK1 ACAGTGTATAAACTCCCACA**TGG**
IKBKE (IKKϵ) TGCATCGCGACATCAAGCCG**GGG**
CHUK (IKKα) TAGTTTAGTAGTAGAACCCA**TGG**
IKBKB (IKKβ) GCCATGGAGTACTGCCAAGG**AGG**
IRF7 CCGAGCTGCACGTTCCTATA**CGG**
IRF3 GTTACTGGGTAACATGGTGT**TGG**
TRIM33 GTTATGAACTTCACAAATT**GGG**
VMP1 GAACTGCCAGTTTGGCCC**GGG**
UBA3 GGCCTAAGGAGCAGCCTTT**TGG**
NAE1 GAATTAAATAGCGATGTCTC**TGG**
NF1 GCTGGTTTCCTTCACGAC**AGG**
USP18 GGCACAGTCAACGCAAATCA**AGG**
NFIC GCTGCTGGGCGAGAAGCCCG**AGG**
EFR3A GATTGCTATGGAGGCAC**TGG**
PTPN1 GAGCAGATCGACAAGTCC**GGG**
MORC3 GCTGATACTGAGATACCATA**TGG**
TAF5L GCTGCTCAATGACATCCTTC**TGG**
RNF7 GGCCATGTGGAGCTGGGACG**TGG**
PI4KA GGGATAGCATACTTGCAA**AGG**
IFNAR1 GTACATTGTATAAAGACCAC**AGG**
IFNAR2 TGAGTGGAGAAGCACACACG**AGG**

STAT1 CAGGAGGTCATGAAAACGGA**TGG**
STAT2 ATCATCTCAGCCAACTGGGT**AGG**
IFNB1 GATGAACTTTGACATCCCTG**AGG**
MRE_1 AACCCTAATGTACACTTGGT**AGG**
MRE_2 CACTTCTAGAACGGTCACCA**TGG**
scramble GCTGCTCCCTAACAGGACGC

5.8.7 Cytokine quantification

Cytokine secretion was quantified by ELISA of cell-free supernatants after stimulations (IFNβ: R&D, DY814-05; IP-10: BD, 550926).

5.8.8 Lenti-/retro-viral transduction

Lenti- and retro-virus was produced in HEK293T cells. 4.5 million cells were plated per 10cm dish and transfected with 5µg of transfer vector, 3.75µg of packaging vector (pd8.9 for lenti- and pGAGPOL for retro-virus) and 1.5µg pVSVG using 30.75µg PEI-MAX (Polysciences, 24765-1). 12h after transfection the medium was replaced with DMEM medium containing 30% (v/v) FCS. After 24h-36h, viral supernatants were harvested, centrifugated at 1000×g for 10min, and filtered through a .45µm filter. Cells were culture for 48 after transduction, prior to selection with puromycin or Blasticidine S hydrochloride (both Sigma-Aldrich).

5.8.9 Ectopic gene expression

A doxycycline-inducible lenti-virus system was used for ectopic gene expression as previously described [59]. Codon-optimized constructs for HSV1-ICP0 and HA-Adenovirus5-E4ORF3 were synthesized by Integrated DNA Technologies and cloned into pLIP. ICP0 variants were generated by overlap-extension PCR.

5.8.10 IFNB1-reporter

The 1kb upstream of the transcription start site of huma IFNB1 (hg38 chr9:21077923-21078922) was synthesized by Integrated DNA Technologies and cloned in front of a luciferase reporter from *Gaussia princeps* [11] into a retro-viral transfer vector in opposite direction to the 5'LTR. BLaER1 cells were transduced and sorted for reporter integration.

5.8.11 Flow Cytometry

BLaER1 were harvested for flow cytometry, fixed and permeabilized using eBioscience™ IC Fixation Buffer and eBioscience™ Permeabilization Buffer (both Thermo Fisher) according to the provider's protocol. Cells were incubated for 1h with PE-Anti-Viperin (Clone MaP.VIP;

BD, 565196) and analyzed using a BD LSRFortessa™ Flow Cytometer. If indicated, cells were sorted on a BD FACSAria™ Fusion Cell Sorter.

5.8.12 CRISPR-Screen

Monoclonal Cas9-expressing BLaER1 cells were re-selected with blasticidine and 9 million cells were transduced in four biological replicates with a pooled Human CRISPR Knockout library at an MOI of approximately 0.3. The library was a gift from Michael Bassik (Addgene #101926, 101927, 101928, 101929, 101930, 101931, 101932, 101933, 101934). Two days after transduction, cells were selected with puromycin for 3 days and trans-differentiated. Cells were stained for Viperin expression. Per biological replicate, $46 - 8010^3$ cells with increased spontaneous Viperin expression were sorted into direct lysis buffer (0.2 mg/ml proteinase K, 1 mM CaCl2, 3 mM MgCl2, 1 mM EDTA, 1% Triton X 100, 10 mM Tris pH 7.5). As control, sgRNA positive cells were sorted irrespectively of their Viperin expression. The reactions were incubated at 65°C for 10min and at 95°C for 15min. The integrated sgRNA cassette was amplified using a nested PCR approach with Phusion DNA Polymerase (Thermo Fisher) with 4 technical replicates per biological replicate.

Details of the nested PCR approach and primers for the second level PCR have been described [183]. PCR products were sequenced on Illumina HiSeq4000 50SR. Deep sequencing data was analyzed with PinAPL-Py [198]. The recovery of the sgRNA library was suboptimal, probably due to low number of sorted cells. Subsequently, reads from all technical and biological replicates were combined and enriched sgRNAs over control were identified. Modified robust ranking aggregation (RRA) to gene level revealed candidate negative regulators of IFN. Candidates with a significantly highly ranked sgRNAs (p.adjust < 0.01) were validated in an arrayed format (Figure 5.1c).

5.8.13 RNA-Seq

RNA from 2.8 million trans-differentiated BLaER1 monocytes was isolated using TRIzol™ Reagent (Thermo Fisher) according to the manufacturer's recommendation. DNA was removed with RQ1 RNase-free DNase (Promega) in the presence of RNasin plus Ribonuclease Inhibitor (Promega) and RNA isolated with Agencourt AMPure XP beads (Beckman Coulter). mRNA-seq libraries were prepared by the QB3 Genomics Functional Genomics Laboratory from poly-A-enriched mRNA using a KAPA mRNA HyperPrep Kit (Roche) and sequenced on the Illumina Novaseq S4 150PE. Sequencing quality of fastq files was evaluated with FASTQC [51] and paired end RNA-seq reads were aligned to the reference genome (GRCh38.83) using Bowtie2 [112] with default settings. Transcript and gene counts were quantified using RSEM [117] with the parameter 'strandedness' set to 'reverse' to account for strand-specific library preparation protocol. DeSeq2 [126] was used to identify differentially expressed genes between conditions by building a single DeSeq2 model using counts from RSEM and performed pairwise comparisons of conditions. We consider all genes with a false discovery rate (FDR) below 0.05 to be significantly differential. We used log normalized

counts from DeSeq2 to perform principal component analysis (PCA) shown in Figure 5.1 and to generate heatmaps. Gene enrichment analysis was done with the R package clusterProfiler [234]. For each enrichment analysis, we assign the foreground to be the set of genes with significant increased expression in the condition being evaluated. We then set the background to the set of all genes that have mean counts greater than 1 for the condition being evaluated. Gene enrichment sets were downloaded from the Molecular Signatures Database (MSigDB) [121].

5.8.14 ATAC-seq

ATAC-seq was performed as previously described by Buenrostro et al. [25]. 50,000 BLaER1 monocytes were washed with PBS and lysed in 50 µL of cold lysis buffer (10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM MgCl2, 0.1% IGEPAL CA-630). Nuclei were harvested by centrifugation and resuspended in 50 µl TD Buffer (Illumina, FC-121-1030) with 2.5 µL Tn5 Transposase (Illumina, FC-121-1030). The reaction was incubated for 30 min at 37°C and DNA was isolated with a MinElute Kit (Qiagen). Transposed DNA fragments were amplified to reach 30% of maximal amplification using Ad1 and Ad2 primer and sequenced on an Illumina Nova-Seq SP 50PE.

ATAC-seq was processed as discussed in Section 4.10.4. We obtained counts representing peak strength for all samples by counting the number of Tn5 cut sites that overlapped each peak for each sample. We used DeSeq2 [126] to identify the differential abundance of cut sites between IFNAR1–/– IFNAR2–/– and MORC3–/–IFNAR1–/– IFNAR2–/– monocytes, and identified all regions with a FDR below 0.05 to be significantly differentially accessible between conditions.

5.8.15 Statistical analysis

Data was analyzed for statistical significant differences using GraphPad Prism [99]. Gene expression values were log2 transformed and viral titers were log10 transformed for statistical analysis. Statistical tests are indicated in the figure legends and were RM one-way or two-way ANOVA with Geisser-Greenhouse correction and Dunnett's post hoc test or paired, two-sided T-test. * $p < 0.05$; ** $p < 0.01$ *** $p < 0.001$.

Part III

Conclusion

Chapter 6

Discussion

In this work, we have introduced a set of tools and methods to aid in the processing and interpretation of epigenetic datasets, with a focus on interpretability of chromatin accessibility. We apply these tools to multiple applications related to innate and adaptive immune systems. Within each chapter, we discuss the implications and future work for each method and application introduced. Therefore, in this remaining chapter, we discuss a subset of broad problems not yet addressed in this work, that should be considered as new methods and tools are developed for interpreting epigenetic datasets.

6.1 Consideration of environmental and user impacts on consumption of computational resources

In the machine learning and data science communities, proving computational efficiency while maintaining or improving upon statistical performance over alternative methods is occasionally used to help convince the broader community that they have made progress over existing methodology [187, 205]. Although numerous methods designed to process and analyze biological datasets have provided information regarding the improvement on runtime over existing methods [124, 145], most of the methods referenced in this paper fail to report metrics of runtimes, or the number of model parameters required to construct a model [166, 170, 5]. Although often overlooked, computational efficiency is an important factor to consider when presenting new methodology. Here, we discuss the environmental and user impact of failing to consider such metrics.

A recent study that evaluated computational resources required to train machine learning models for natural language processing (NLP) demonstrated that state-of-the-art models produced up to 626,000 pounds of CO_2 emissions for training a single model, with minimal improvement in accuracy over existing methods [202]. From these findings, the authors concluded that proposed methods should provide metrics that allow for direct comparison to existing state-of-the-art, including training time and information about hyperparameters. Recent studies have proposed more concrete sets of metrics to evaluate computational ef-

ficiency, including carbon emission, runtime, and the number of floating point operations required to generate results [187]. Regardless, providing such information would allow users to make educated decisions of model selection, based on their own cost and computational constraints.

Ignorance of the consideration of resources required to train and validate models can inhibit individual researchers that do not have available compute resources to reproduce results. Lack of consideration of runtime and required resources during model conceptualization may prove the model unusable. One such example in which model design hinders usage is in the application of Epitome (Section 2.3.1) to scATAC-seq datasets (Section 2.3.2). Although Epitome models train quickly on generic computer hardware [144], applying trained models to predict binding over multiple cellular conditions is sequential, rendering the application of Epitome models to predict TFBS in tens of thousands of cells slow (30 hours for a scATAC-seq dataset with 10,000 cells over 130,000 genomic regions). Methods such as chromVAR [181] forgo resolution of TFBS at individual genomic loci by directly computing cell-level metrics of TF binding in a single computation that takes less than two hours [30]. Thus, tradeoffs such as accuracy, resolution, and runtime can be directly compared to justify certain design decisions during model conceptualization.

6.2 Towards open source and reproducible projects

When trying to prove advancements of a computational method, much of our time is spent learning how to configure and reproduce results of state-of-the-art methods apart from our own. Although journals have recently made strides to require availability of code that is central to the findings of a publication [33], missing or incomplete documentation and unspecified hardware requirements still make reproducing results difficult [163]. Both Epitome [144] and Mango [145] follow robust protocol that aids in support for open-source reproducible software. These software implement documentation, unit and integration tests, as well as code checking standards. Through the Computational Biology Skills Seminar at University of California, Berkeley, we have provided tutorials and resources to aid researchers in the development of reproducible python packages.

Acronyms

API Application Programming Interface.

bp base pair.

CKS convolutional kitchen sink.

CNN convolutional neural network.

ERV Endogenous retrovirus.

FDR false discovery rate.

HSC hematopoietic stem cell.

HSV-1 Herpes Simplex Virus-1.

IFN interferon.

ISG interferon-stimulated gene.

KO knock out.

MACS magnetic-activated cell sorting.

mRNA messanger ribonucleic acid.

PAMP pathogen-associated-molecular patterns.

PBMC peripheral blood mononuclear cell.

PIC pre-initialization complex.

PRR pattern-recognition-receptor.

PTM post-translational modification.

ROC Receiver operating characteristic.

STING stimulator of interferon genes.

TF transcription factor.

TFBS transcription factor binding sites.

Viperin Virus inhibitory protein, endoplasmic reticulum-associated, interferon-inducible.

WT wild type.

Appendix A

Get a Hobby

Getting a PhD is all about becoming an expert in one area. However, this doesn't mean you can't excel in areas outside of your research during grad school. During my time at UC Berkeley, I transformed from an amateur biker into a skilled rider. Here, I share a couple of my favorite bike rides around Berkeley.

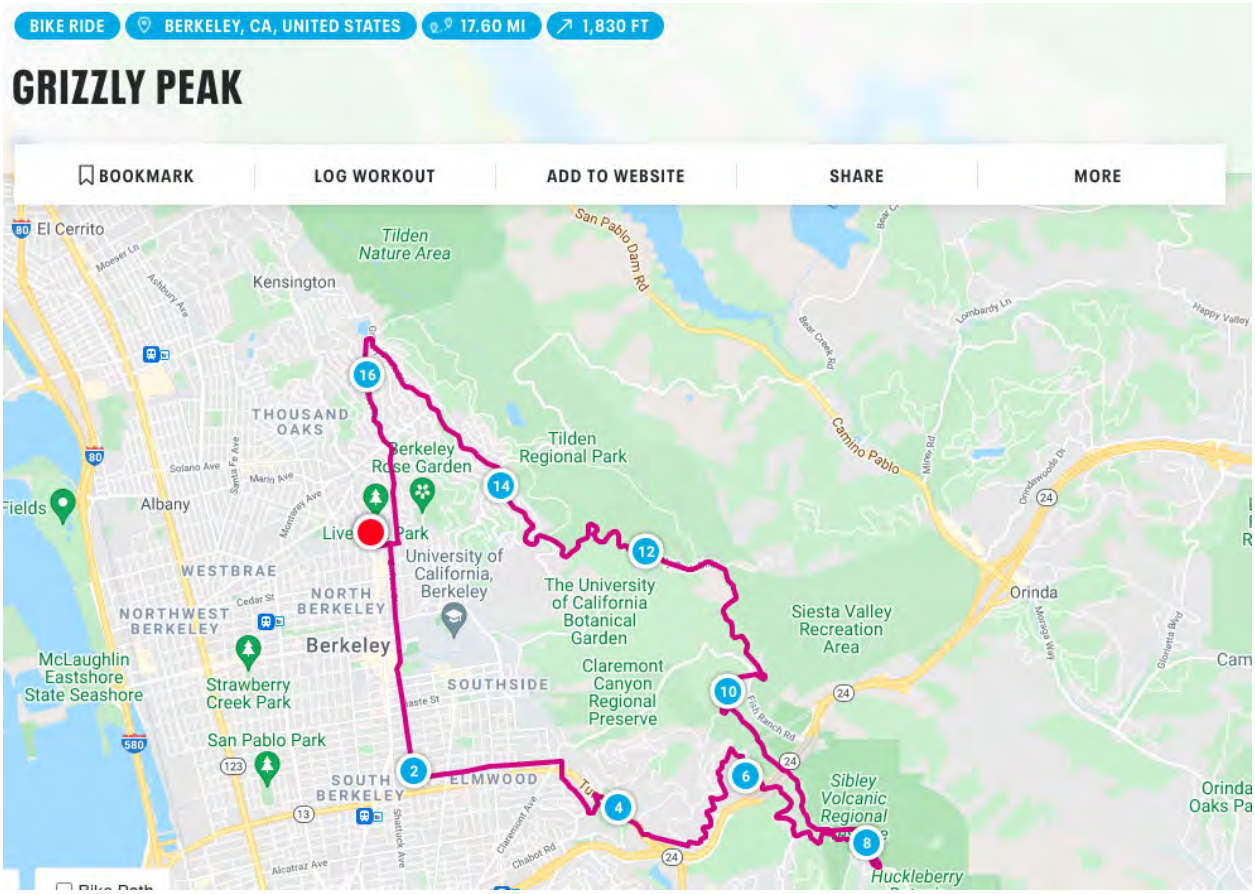


Figure A.1: Grizzly peak is a great weekday ride: it is only 17 miles, and can be completed before or after going in to lab. This ride includes a good ascent for training, but not too much to kill you.

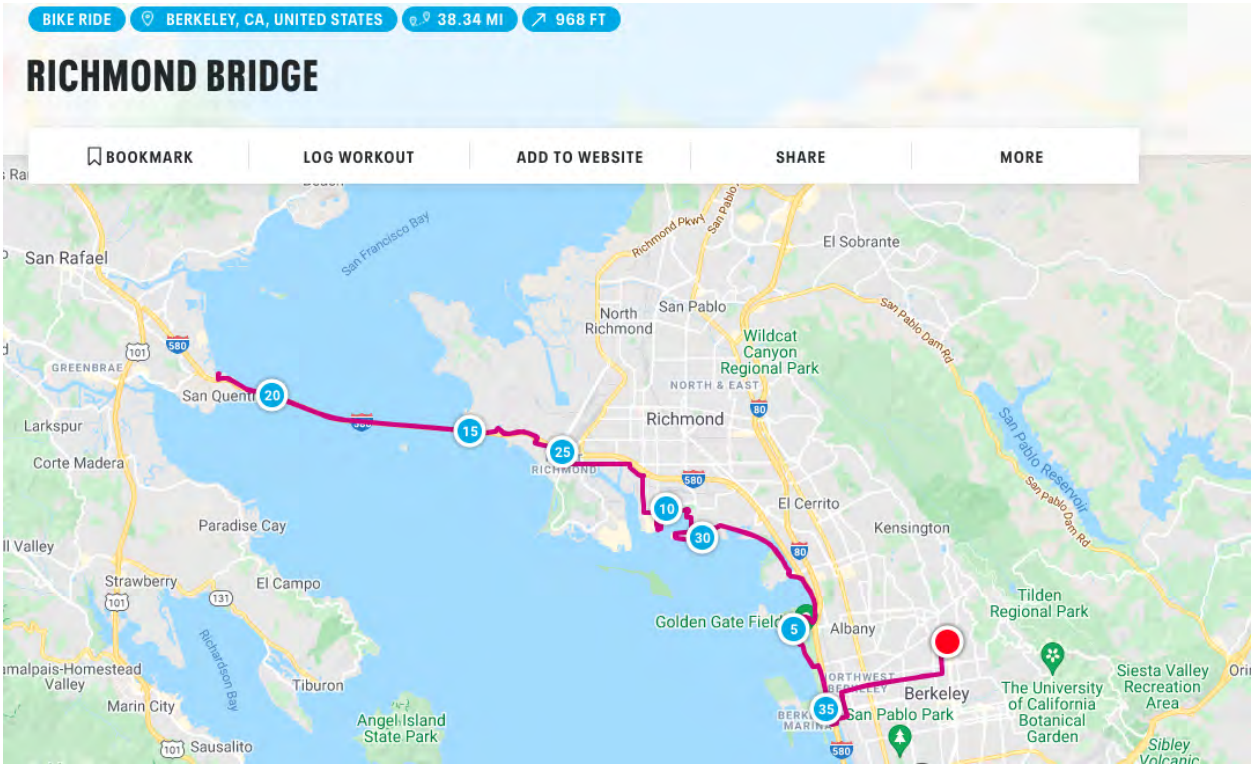


Figure A.2: Biking the Richmond bridge is a great way to get out of the city and see the water. It is relatively flat, passes pretty marinas, and leads to a great view of the bay. This is a feel-good ride.

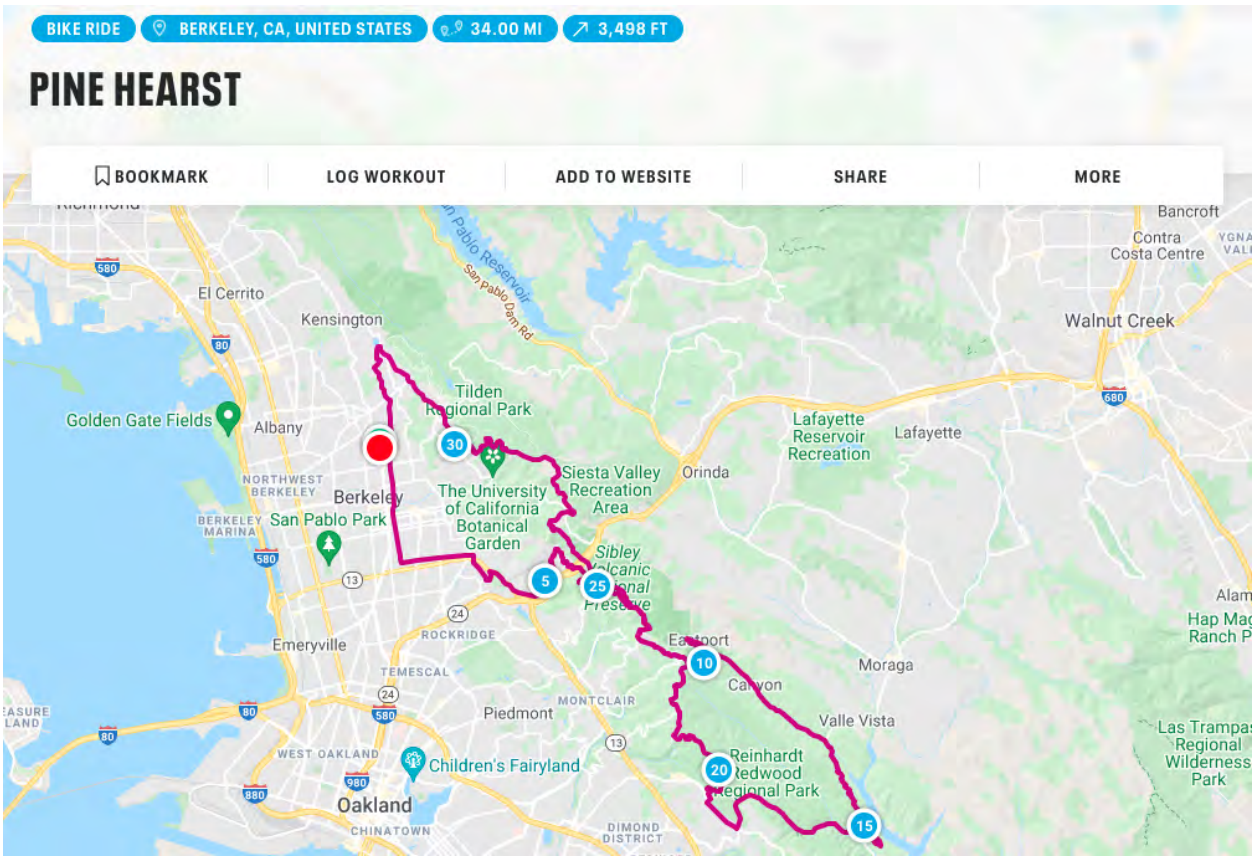


Figure A.3: Pine Hearst is a great ride to do with friends on the weekend. It is nice on hotter days, as much of it is shaded by the redwoods. Much of it is rolling hills so you never get bored!

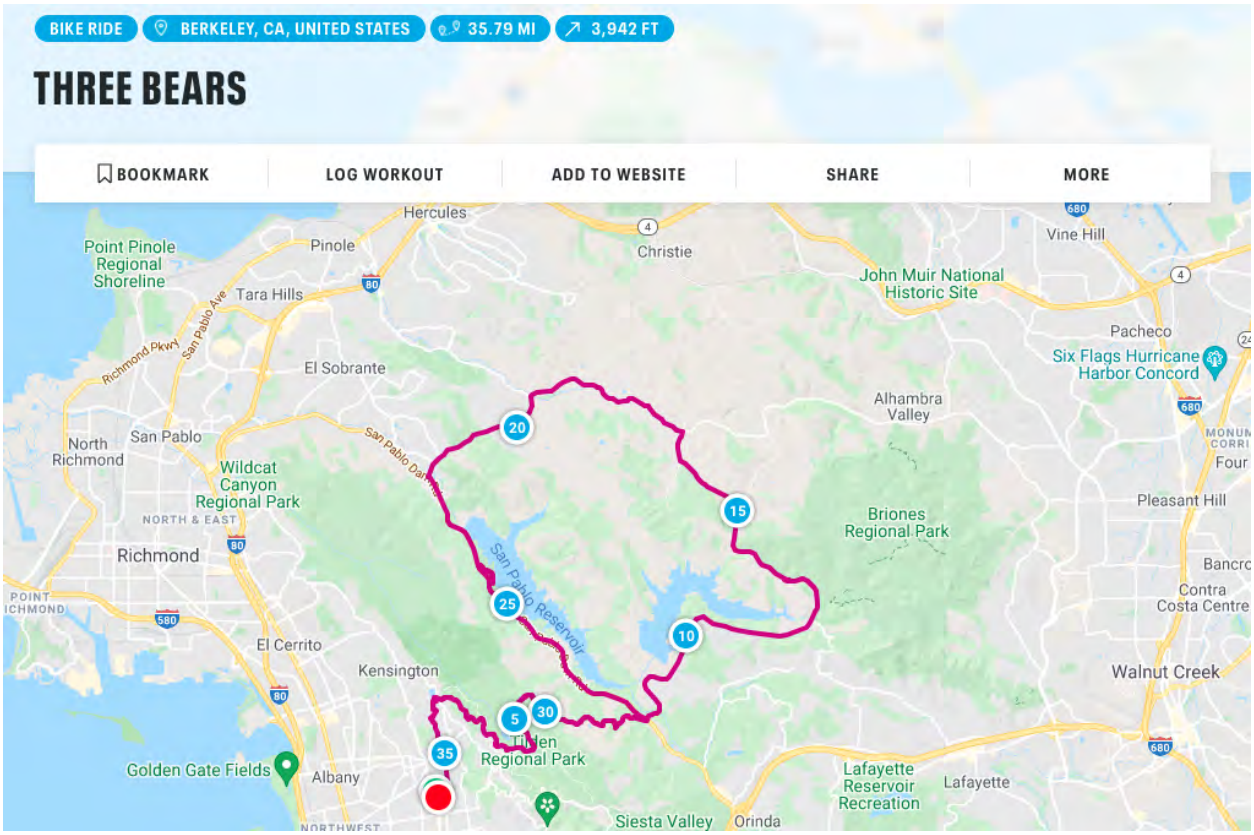


Figure A.4: Three Bears is another weekend favorite. On this ride, you power through three main ascents: the papa, mama, and baby. Keep an eye out for horses and cows.

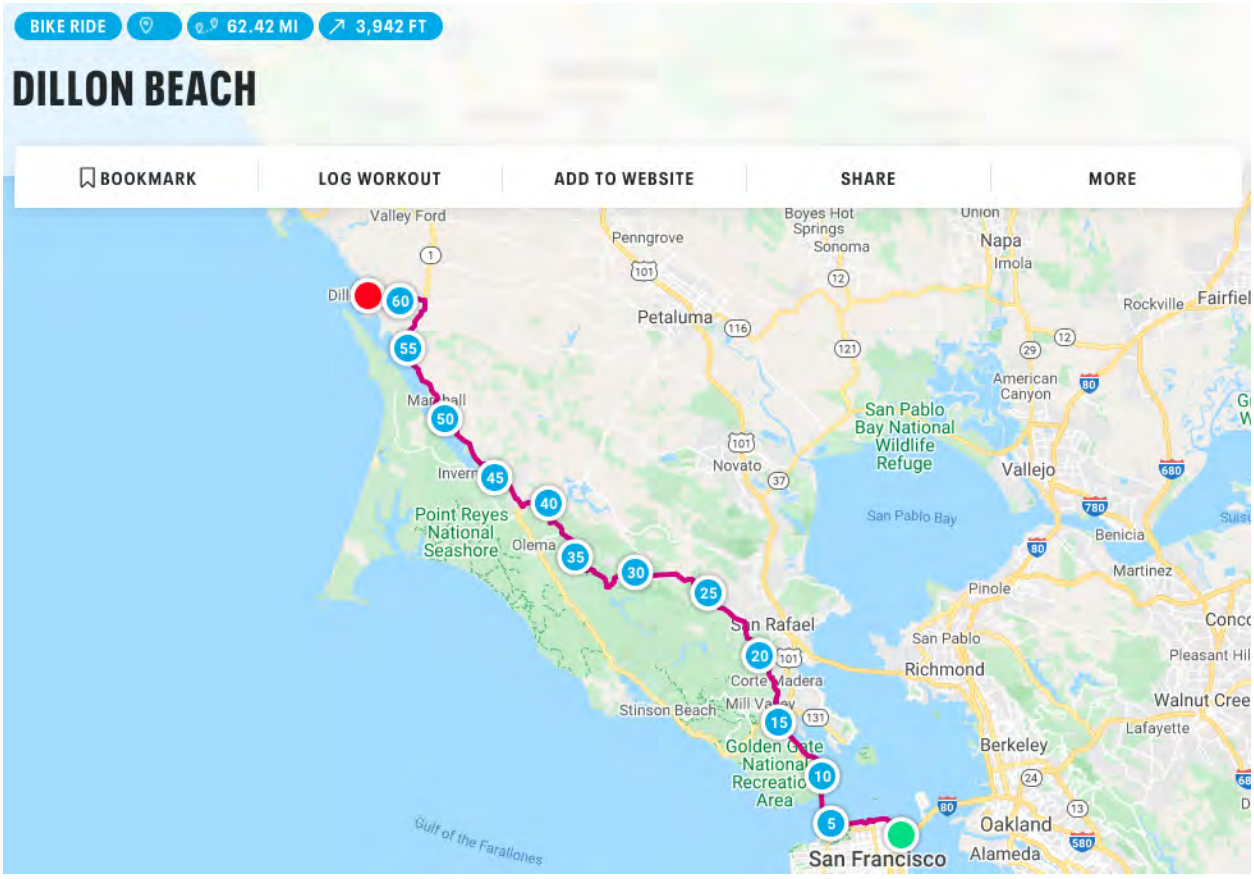


Figure A.5: If you are looking for an overnight bike ride, pack up your toothbrush and swimsuit and bike to Dillon Beach. This ride passes by numerous cheese companies in Marin, so you can stop by for a tasting.

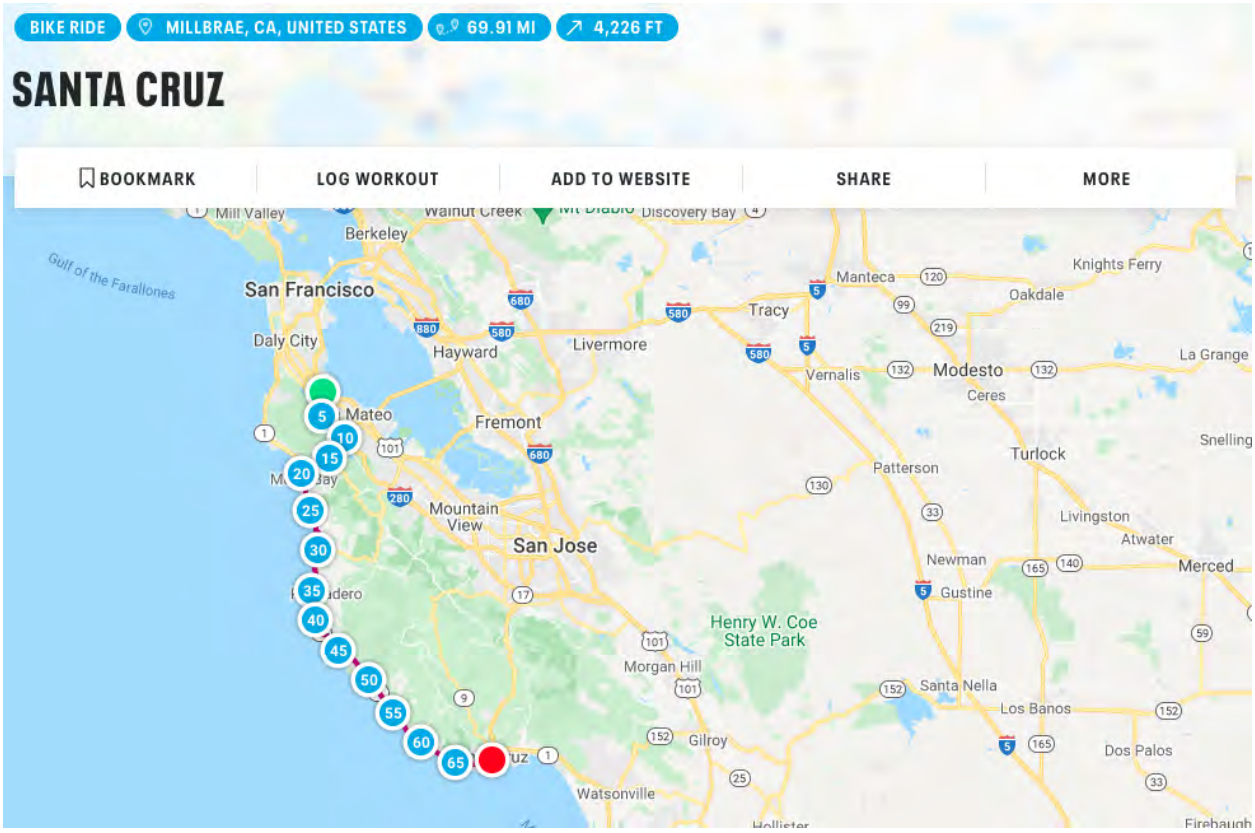


Figure A.6: This is yet another great overnight bike ride. You can find a place to stay in Santa Cruz and relax by the beach after a long ride down Hwy 1.

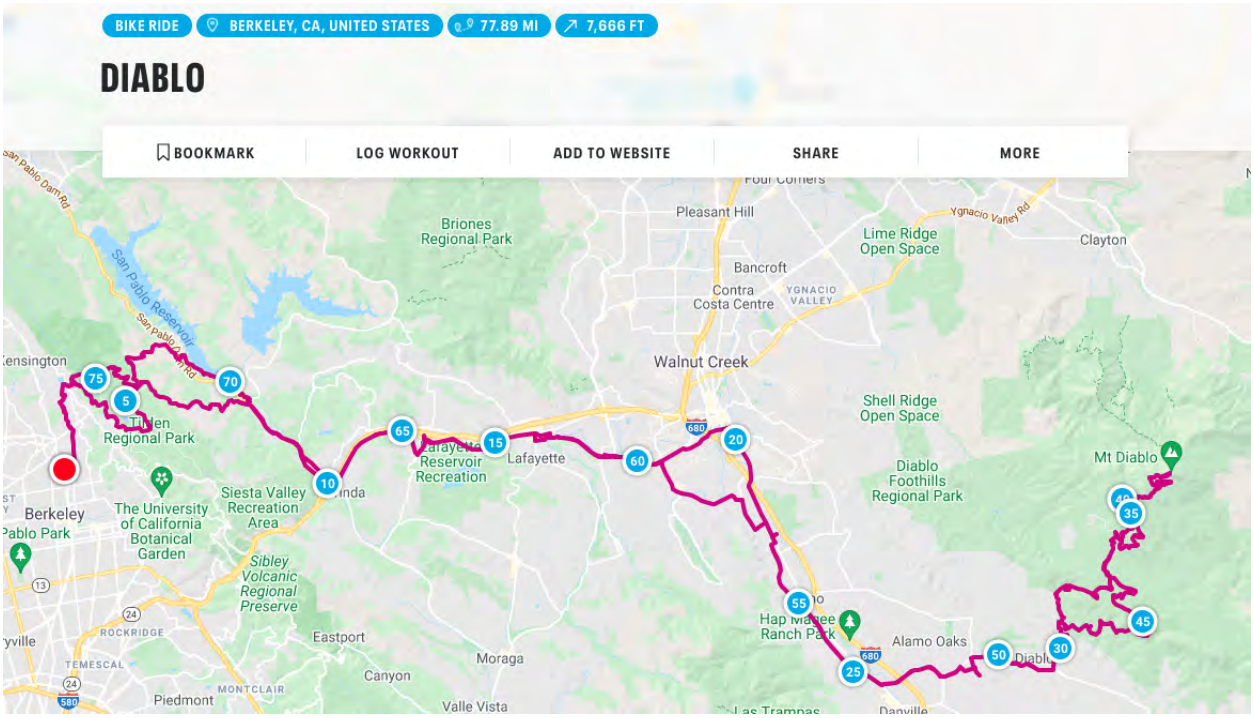


Figure A.7: With over 7,000 feet of ascent, Mount Diablo is a great challenge. Make sure to do this one in the cooler months.

Bibliography

[1] 1000 Genomes Project Consortium et al. "A global reference for human genetic variation." In: *Nature* 526.7571 (Oct. 2015), pp. 68–74. ISSN: 1476-4687. DOI: 10.1038/nature15393. URL: http://dx.doi.org/10.1038/nature15393.

[2] Shizuo Akira, Satoshi Uematsu, and Osamu Takeuchi. "Pathogen recognition and innate immunity." eng. In: *Cell* 124.4 (Feb. 24, 2006), pp. 783–801. ISSN: 0092-8674 (Print); 0092-8674 (Linking). DOI: 10.1016/j.cell.2006.02.015.

[3] Cristina M. Alberini and Eric Klann. "Chapter 5 - Regulation of Neuronal Gene Expression and Protein Synthesis". In: *From Molecules to Networks (Third Edition)*. Ed. by John H. Byrne, Ruth Heidelberger, and M. Neal Waxham. Third Edition. Boston: Academic Press, 2014, pp. 149–174. DOI: https://doi.org/10.1016/B978-0-12-397179-1.00005-1. URL: https://www.sciencedirect.com/science/article/pii/B9780123971791000051.

[4] Alex Max Monteys and Amiel Hundley. "Regulated control of gene therapies by drug-induced splicing". In: *Nature* (June 2021).

[5] Babak Alipanahi et al. "Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning". In: *Nature biotechnology* (2015).

[6] C. Alvarez-Fernández et al. "A short CD3/CD28 costimulation combined with IL-21 enhance the generation of human memory stem T cells for adoptive immunotherapy." eng. In: *Journal of translational medicine* 14.1 (July 2016), p. 214. ISSN: 1479-5876. DOI: 10.1186/s12967-016-0973-y.

[7] Takashi Aoi et al. "Generation of pluripotent stem cells from adult mouse liver and stomach cells." eng. In: *Science (New York, N.Y.)* 321.5889 (Aug. 2008). Place: United States, pp. 699–702. ISSN: 1095-9203 0036-8075. DOI: 10.1126/science.1154884.

[8] Tal Ashuach et al. "PeakVI: A Deep Generative Model for Single Cell Chromatin Accessibility Analysis". In: *bioRxiv* (2021). DOI: 10.1101/2021.04.29.442020. eprint: https://www.biorxiv.org/content/early/2021/04/30/2021.04.29.442020.full.pdf. URL: https://www.biorxiv.org/content/early/2021/04/30/2021.04.29.442020.

[9] Regina Bailey. "What is Chromatin's Structure and Function?" In: *ThoughtCo* (Aug. 2020).

[10] Andrew J Bannister and Tony Kouzarides. "Regulation of chromatin by histone modifications". In: *Cell Research* 21.3 (2011), pp. 381–395. DOI: 10.1038/cr.2011.22. URL: https://doi.org/10.1038/cr.2011.22.

[11] Eva Bartok et al. "iGLuc: a luciferase-based inflammasome and protease activity reporter". In: *Nature Methods* 10.2 (Jan. 2013), pp. 147–154. DOI: 10.1038/nmeth.2327. URL: https://doi.org/10.1038%2Fnmeth.2327.

[12] Erik A. Van Der Biezen and Jonathan D.G. Jones. "Plant disease-resistance proteins and the gene-for-gene concept". In: *Trends in Biochemical Sciences* 23.12 (Dec. 1998), pp. 454–456. DOI: 10.1016/s0968-0004(98)01311-5. URL: https://doi.org/10.1016%2Fs0968-0004%2898%2901311-5.

[13] Alexander Birbrair and Paul S Frenette. "Niche heterogeneity in the bone marrow." eng. In: *Ann N Y Acad Sci* 1370.1 (Apr. 2016), pp. 82–96. ISSN: 1749-6632 (Electronic); 0077-8923 (Print); 0077-8923 (Linking). DOI: 10.1111/nyas.13016.

[14] Michael E. Birnbaum et al. "Molecular architecture of the $\alpha\beta$ T cell receptor–CD3 complex". In: *Proceedings of the National Academy of Sciences* 111.49 (2014), pp. 17576–17581. ISSN: 0027-8424. DOI: 10.1073/pnas.1420936111. URL: https://www.pnas.org/content/111/49/17576.

[15] John A Blaho, Elise R Morton, and Jamie C Yedowitz. "Herpes simplex virus: propagation, quantification, and storage." eng. In: *Curr Protoc Microbiol* Chapter 14 (Oct. 10, 2005), Unit 14E.1. ISSN: 1934-8533 (Electronic); 1934-8525 (Linking). DOI: 10.1002/9780471729259.mc14e01s00.

[16] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. "Trimmomatic: a flexible trimmer for Illumina sequence data". In: *Bioinformatics* 30.15 (Aug. 2014), pp. 2114–2120. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu170. URL: https://doi.org/10.1093/bioinformatics/btu170 (visited on 05/18/2021).

[17] Chris Boutell et al. "A Viral Ubiquitin Ligase Has Substrate Preferential SUMO Targeted Ubiquitin Ligase Activity that Counteracts Intrinsic Antiviral Defence". In: *PLoS Pathogens* 7.9 (Sept. 2011). Ed. by Karen L. Mossman, e1002245. DOI: 10.1371/journal.ppat.1002245. URL: https://doi.org/10.1371%2Fjournal.ppat.1002245.

[18] Eric D Boyden and William F Dietrich. "Nalp1b controls mouse macrophage susceptibility to anthrax lethal toxin". In: *Nature Genetics* 38.2 (Jan. 2006), pp. 240–244. DOI: 10.1038/ng1724. URL: https://doi.org/10.1038%2Fng1724.

[19] Alan P Boyle et al. "High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells." eng. In: *Genome Res* 21.3 (Mar. 1, 2011), pp. 456–464. ISSN: 1549-5469 (Electronic); 1088-9051 (Print); 1088-9051 (Linking). DOI: 10.1101/gr.112656.110.

[20] Alan P Boyle et al. "High-resolution mapping and characterization of open chromatin across the genome". In: *Cell* 132.2 (Jan. 2008), pp. 311–322.

[21] David K Breslow et al. "Orm family proteins mediate sphingolipid homeostasis." In: *Nature* 463.7284 (Feb. 2010). Publisher: Springer Nature, pp. 1048–53. DOI: 10.1038/NATURE08787. URL: http://www.ncbi.nlm.nih.gov/pubmed/20182505.

[22] Jennifer N. Brudno et al. "T Cells Genetically Modified to Express an Anti-B-Cell Maturation Antigen Chimeric Antigen Receptor Cause Remissions of Poor-Prognosis Relapsed Multiple Myeloma." eng. In: *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 36.22 (Aug. 2018), pp. 2267–2280. ISSN: 1527-7755 0732-183X. DOI: 10.1200/JCO.2018.77.8084.

[23] P. van der Bruggen et al. "A gene encoding an antigen recognized by cytolytic T lymphocytes on a human melanoma." eng. In: *Science (New York, N.Y.)* 254.5038 (Dec. 1991). Place: United States, pp. 1643–1647. ISSN: 0036-8075. DOI: 10.1126/science.1840703.

[24] Jason D Buenrostro et al. "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position". In: *Nature methods* 10.12 (2013), pp. 1213–1218.

[25] Jason D. Buenrostro et al. "ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide." eng. In: *Current protocols in molecular biology* 109 (Jan. 2015), pp. 21.29.1–21.29.9. ISSN: 1934-3647 1934-3639. DOI: 10.1002/0471142727.mb2129s109.

[26] Jason D. Buenrostro et al. "Single-cell chromatin accessibility reveals principles of regulatory variation". In: *Nature* 523.7561 (2015), pp. 486–490. DOI: 10.1038/nature14590. URL: https://doi.org/10.1038/nature14590.

[27] Chih-Chung Chang and Chih-Jen Lin. "LIBSVM: A library for support vector machines". In: *ACM Transactions on Intelligent Systems and Technology* 2.27 (3 2011). DOI: 10.1145/1961189.1961199.

[28] Francisco Charte et al. "MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation". In: *Knowl. Based Syst.* 89 (2015), pp. 385–397.

[29] Ailing Chen, Daozhen Chen, and Ying Chen. "Advances of DNase-seq for mapping active gene regulatory elements across the genome in animals". In: *Gene* 667 (2018), pp. 83–94.

[30] Huidong Chen et al. "Assessment of computational methods for the analysis of single-cell ATAC-seq data". In: *Genome Biology* 20.1 (2019), p. 241. DOI: 10.1186/s13059-019-1854-5. URL: https://doi.org/10.1186/s13059-019-1854-5.

[31] Xiaoyu Chen et al. "A dynamic Bayesian network for identifying protein-binding footprints from single molecule-based sequencing data." eng. In: *Bioinformatics* 26.12 (June 15, 2010), pp. i334–42. ISSN: 1367-4811 (Electronic); 1367-4803 (Print); 1367-4803 (Linking). DOI: 10.1093/bioinformatics/btq175.

[32] Paula M Chilton and Thomas C Mitchell. "CD8 T cells require Bcl-3 for maximal gamma interferon production upon secondary exposure to antigen." In: *Infection and Immunity* (2005). DOI: 10.1128/IAI.01749-05. URL: https://doi.org/10.1128/IAI.01749-05.

[33] "Code share". In: *Nature* 514.7524 (2014), pp. 536–536. DOI: 10.1038/514536a. URL: https://doi.org/10.1038/514536a.

[34] ENCODE Project Consortium et al. "The ENCODE (ENCyclopedia of DNA elements) project". In: *Science* 306.5696 (2004), pp. 636–640.

[35] The ENCODE Project Consortium. "An Integrated Encyclopedia of DNA Elements in the Human Genome". In: *Nature* (2012), pp. 57–74. DOI: 10.1038/nature11247.

[36] Richard Cordaux and Mark A Batzer. "The impact of retrotransposons on human genome evolution." eng. In: *Nat Rev Genet* 10.10 (Oct. 9, 2009), pp. 691–703. ISSN: 1471-0064 (Electronic); 1471-0056 (Print); 1471-0056 (Linking). DOI: 10.1038/nrg2640.

[37] Steven W. Criscione et al. "Transcriptional landscape of repetitive elements in normal and cancer human cells". In: *BMC Genomics* 15.1 (2014), p. 583. DOI: 10.1186/1471-2164-15-583. URL: https://doi.org/10.1186/1471-2164-15-583.

[38] Darren A Cusanovich et al. "Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing." eng. In: *Science* 348.6237 (May 22, 2015), pp. 910–914. ISSN: 1095-9203 (Electronic); 0036-8075 (Print); 0036-8075 (Linking). DOI: 10.1126/science.aab1601.

[39] Curt A Davey et al. "Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 a resolution." eng. In: *J Mol Biol* 319.5 (June 21, 2002), pp. 1097–1113. ISSN: 0022-2836 (Print); 0022-2836 (Linking). DOI: 10.1016/S0022-2836(02)00386-8.

[40] David DeTomaso et al. "Functional interpretation of single cell similarity maps". In: *Nature Communications* 10.1 (2019), p. 4376. DOI: 10.1038/s41467-019-12235-0. URL: https://doi.org/10.1038/s41467-019-12235-0.

[41] Niall Dillon. "Gene regulation and large-scale chromatin organization in the nucleus". In: *Chromosome Research* (2006), pp. 117–126. URL: https://link.springer.com/article/10.1007/s10577-006-1027-8.

[42] Alexander Dobin et al. "STAR: ultrafast universal RNA-seq aligner." eng. In: *Bioinformatics* 29.1 (Jan. 1, 2013), pp. 15–21. ISSN: 1367-4811 (Electronic); 1367-4803 (Print); 1367-4803 (Linking). DOI: 10.1093/bioinformatics/bts635.

[43] Paula Dobosz and Tomasz Dzieciatkowski. "The Intriguing History of Cancer Immunotherapy." eng. In: *Frontiers in immunology* 10 (2019), p. 2965. ISSN: 1664-3224. DOI: 10.3389/fimmu.2019.02965.

[44] I. Dunhan, A. Kundaje, and S. et al. Aldred. "An integrated encyclopedia of DNA elements in the human genome." In: *Nature* 489 (2012), pp. 57–74.

[45] Timothy J. Durham et al. "PREDICTD PaRallel Epigenomics Data Imputation with Cloud-based Tensor Decomposition". In: *Nature Communications* 9.1 (2018), p. 1402.

[46] Shimon Efrat. "Epigenetic Memory: Lessons From iPS Cells Derived From Human β Cells". In: *Frontiers in Endocrinology* 11 (2021), p. 1063. ISSN: 1664-2392. DOI: 10.3389/fendo.2020.614234. URL: https://www.frontiersin.org/article/10.3389/fendo.2020.614234.

[47] Mohamed A ElTanbouly et al. "VISTA is a checkpoint regulator for naıve T cell quiescence and peripheral tolerance." eng. In: *Science* 367.6475 (Jan. 17, 2020). ISSN: 1095-9203 (Electronic); 0036-8075 (Print); 0036-8075 (Linking). DOI: 10.1126/science.aay0524.

[48] Jason Ernst and Manolis Kellis. "Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues". In: *Nature Biotechnology* 33.4 (2015), pp. 364–376.

[49] Z. Eshhar. "The T-body approach: redirecting T cells with antibody specificity." eng. In: *Handbook of experimental pharmacology* 181 (2008). Place: Germany, pp. 329–342. ISSN: 0171-2004. DOI: 10.1007/978-3-540-73259-4_14.

[50] Eleazar Eskin et al. "Mismatch string kernels for SVM protein classification". In: *Advances in Neural Information Processing Systems*. 2002, pp. 1417–1424.

[51] *FastQC*. June 2015. URL: https://qubeshub.org/resources/fastqc.

[52] Natasha Lopes Fischer et al. "Effector-triggered immunity and pathogen sensing in metazoans". In: *Nature Microbiology* 5.1 (Dec. 2019), pp. 14–26. DOI: 10.1038/s41564-019-0623-2. URL: https://doi.org/10.1038%2Fs41564-019-0623-2.

[53] Marc Fiume et al. "Savant: Genome Browser for High Throughput Sequencing Data". In: *Bioinformatics (Oxford, England)* 26 (Aug. 2010), pp. 1938–44. DOI: 10.1093/bioinformatics/btq332.

[54] Christopher Fletez-Brant et al. "Kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic data sets". In: *Nucleic acids research* 41.W1 (2013), W544–W556.

[55] Oriol Fornes et al. "JASPAR 2020: update of the open-access database of transcription factor binding profiles". In: *Nucleic Acids Research* 48.D1 (Nov. 2019), pp. D87–D92. ISSN: 0305-1048. DOI: 10.1093/nar/gkz1001. eprint: https://academic.oup.com/nar/article-pdf/48/D1/D87/31697272/gkz1001_supplemental_file.pdf. URL: https://doi.org/10.1093/nar/gkz1001.

[56] H. N. Freese, C. D. Norris, and E. A. Loraine. "Integrated Genome Browser: Visual analytics platform for genomics". In: *Bioinformatics* 32.14 (2016), pp. 2089–95.

[57] Yasumitsu Fujie et al. "New type of Sendai virus vector provides transgene-free iPS cells derived from chimpanzee blood." eng. In: *PloS one* 9.12 (2014), e113052. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0113052.

[58] Moritz M. Gaidt et al. "Human Monocytes Engage an Alternative Inflammasome Pathway". In: *Immunity* 44.4 (Apr. 2016), pp. 833–846. DOI: 10.1016/j.immuni.2016.01.012. URL: https://doi.org/10.1016%2Fj.immuni.2016.01.012.

[59] Moritz M. Gaidt et al. "The DNA Inflammasome in Human Myeloid Cells Is Initiated by a STING-Cell Death Program Upstream of NLRP3". In: *Cell* 171.5 (Nov. 2017), 1110–1124.e18. DOI: 10.1016/j.cell.2017.09.039. URL: https://doi.org/10.1016%2Fj.cell.2017.09.039.

[60] Luca Gattinoni, Christopher A Klebanoff, and Nicholas P Restifo. "Paths to stemness: building the ultimate antitumour T cell." In: *Nature Reviews Cancer* 12.10 (Sept. 2012). Publisher: Springer Nature, pp. 671–84. DOI: 10.1038/NRC3322. URL: http://www.ncbi.nlm.nih.gov/pubmed/22996603.

[61] Luca Gattinoni et al. "A human memory T cell subset with stem cell-like properties". eng. In: *Nature medicine* 17.10 (Sept. 2011), pp. 1290–1297. ISSN: 1546-170X. DOI: 10.1038/nm.2446. URL: https://pubmed.ncbi.nlm.nih.gov/21926977.

[62] Luca Gattinoni et al. "T memory stem cells in health and disease." eng. In: *Nature medicine* 23.1 (Jan. 2017), pp. 18–27. ISSN: 1546-170X 1078-8956. DOI: 10.1038/nm.4241.

[63] Mahmoud Ghandi et al. "Enhanced regulatory sequence prediction using gapped k-mer features". In: *PLOS Computational Bioliology* 10.7 (2014), e1003711.

[64] Saba Ghassemi et al. "Reducing Ex Vivo Culture Improves the Antileukemic Activity of Chimeric Antigen Receptor (CAR) T Cells." eng. In: *Cancer immunology research* 6.9 (Sept. 2018), pp. 1100–1109. ISSN: 2326-6074 2326-6066. DOI: 10.1158/2326-6066.CIR-17-0405.

[65] Matan Goldshtein et al. "Transcription Factor Binding in Embryonic Stem Cells Is Constrained by DNA Sequence Repeat Symmetry". In: *Biophysical Journal* 118.8 (2020), pp. 2015–2026.

[66] Sophia Groh and Gunnar Schotta. "Silencing of endogenous retroviruses by heterochromatin." eng. In: *Cell Mol Life Sci* 74.11 (June 2, 2017), pp. 2055–2065. ISSN: 1420-9071 (Electronic); 1420-682X (Linking). DOI: 10.1007/s00018-017-2454-8.

[67] Sophia Groh et al. "Morc3 silences endogenous retroviruses by enabling Daxx-mediated H3.3 incorporation". In: *bioRxiv* (2020). DOI: 10.1101/2020.11.12.380204. eprint: https://www.biorxiv.org/content/early/2020/11/12/2020.11.12.380204.full.pdf. URL: https://www.biorxiv.org/content/early/2020/11/12/2020.11.12.380204.

[68] Kevin Grosselin et al. "High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer". In: *Nature Genetics* 51.6 (2019), pp. 1060–1066. DOI: 10.1038/s41588-019-0424-9. URL: https://doi.org/10.1038/s41588-019-0424-9.

[69] Z Gu. *rGREAT: Client for GREAT Analysis.* 2020. URL: https://github.com/jokergoo/rGREAT,%20http://great.stanford.edu/public/html/.

[70] Eduardo G Gusmao et al. "Detection of active transcription factor binding sites with the combination of DNase hypersensitivity and histone modifications." eng. In: *Bioinformatics* 30.22 (Nov. 15, 2014), pp. 3143–3151. ISSN: 1367-4811 (Electronic); 1367-4803 (Linking). DOI: 10.1093/bioinformatics/btu519.

[71] Jacob Hanna et al. "Direct Reprogramming of Terminally Differentiated Mature B Lymphocytes to Pluripotency". In: *Cell* 133.2 (Apr. 2008), pp. 250–264. ISSN: 0092-8674. DOI: 10.1016/j.cell.2008.03.028. URL: https://www.sciencedirect.com/science/article/pii/S0092867408004479.

[72] Charles R. Harris et al. "Array programming with NumPy". In: *Nature* 585.7825 (Sept. 2020), pp. 357–362. DOI: 10.1038/s41586-020-2649-2. URL: https://doi.org/10.1038/s41586-020-2649-2.

[73] Sergiu Hart. *Shapley Value. In: Game Theory.* Reading, Massachusetts: Palgrave Macmillan, London, 1989, pp. 210–216.

[74] Thomas Hennig et al. "HSV-1-induced disruption of transcription termination resembles a cellular stress response but selectively increases chromatin accessibility downstream of genes." eng. In: *PLoS pathogens* 14.3 (Mar. 2018), e1006954. ISSN: 1553-7374 1553-7366. DOI: 10.1371/journal.ppat.1006954.

[75] Jay R Hesselberth et al. "Global mapping of protein-DNA interactions in vivo by digital genomic footprinting." eng. In: *Nat Methods* 6.4 (Apr. 23, 2009), pp. 283–289. ISSN: 1548-7105 (Electronic); 1548-7091 (Print); 1548-7091 (Linking). DOI: 10.1038/nmeth.1313.

[76] Christian S. Hinrichs et al. "HPV-targeted tumor-infiltrating lymphocytes for cervical cancer." In: *Journal of Clinical Oncology* 32.18_suppl (2014). _eprint: https://doi.org/10.1200/jco.201 LBA3008–LBA3008. DOI: 10.1200/jco.2014.32.18_suppl.lba3008. URL: https://doi.org/10.1200/jco.2014.32.18_suppl.lba3008.

[77] Daniel Quang Hongyang Li and Yuanfang Guan. "Anchor: trans-cell type prediction of transcription factor binding sites". In: *Genome Res.* 29 (2019), pp. 281–292.

[78] Karl-Peter Hopfner and Veit Hornung. "Molecular mechanisms and cellular functions of cGAS–STING signalling". In: *Nature Reviews Molecular Cell Biology* 21.9 (May 2020), pp. 501–521. DOI: 10.1038/s41580-020-0244-x. URL: https://doi.org/10.1038%2Fs41580-020-0244-x.

[79] "How eukaryotic genes are transcribed". In: *Critical reviews in biochemistry and molecular biology* (2009), pp. 117–41. URL: https://pubmed.ncbi.nlm.nih.gov/19514890/.

[80] W. Huber et al. "Orchestrating high-throughput genomic analysis with Bioconductor". In: *Nature Methods* 12 (Jan. 2015). Perspective. URL: https://doi.org/10.1038/nmeth.3252.

[81] J. D. Hunter. "Matplotlib: A 2D Graphics Environment". In: *Computing in Science Engineering* 9.3 (May 2007), pp. 90–95. ISSN: 1521-9615. DOI: 10.1109/MCSE.2007.55.

[82] igvteam. "igv.js". In: (2015). URL: https://github.com/igvteam/igv.js.

[83] Fumitaka Inoue et al. "Identification and Massively Parallel Characterization of Regulatory Elements Driving Neural Induction." In: *Cell Stem Cell* 25.5 (Oct. 2019). Publisher: Elsevier, 713–727.e10. DOI: PMID:1707922. URL: http://www.ncbi.nlm.nih.gov/pubmed/31631012.

[84] Shoichi Iriguchi et al. "A clinically applicable and scalable method to regenerate T-cells from iPSCs for off-the-shelf T-cell immunotherapy". In: *Nature Communications* 12.1 (Jan. 2021), p. 430. ISSN: 2041-1723. DOI: 10.1038/s41467-020-20658-3. URL: https://doi.org/10.1038/s41467-020-20658-3.

[85] Hiroki Ishikawa, Zhe Ma, and Glen N. Barber. "STING regulates intracellular DNA-mediated, type I interferon-dependent innate immunity". In: *Nature* 461.7265 (Sept. 2009), pp. 788–792. DOI: 10.1038/nature08476. URL: https://doi.org/10.1038%2Fnature08476.

[86] Wakana Iwasaki et al. "Contribution of histone N-terminal tails to the structure and stability of nucleosomes." eng. In: *FEBS Open Bio* 3 (2013), pp. 363–369. ISSN: 2211-5463 (Print); 2211-5463 (Electronic); 2211-5463 (Linking). DOI: 10.1016/j.fob.2013.08.007.

[87] Tommi S Jaakkola, Mark Diekhans, and David Haussler. "Using the Fisher kernel method to detect remote protein homologies." In: *Proceedings of ISMB*. Vol. 99. 1999, pp. 149–158.

[88] C.A. Janeway. "Approaching the Asymptote? Evolution and Revolution in Immunology". In: *Cold Spring Harbor Symposia on Quantitative Biology* 54.0 (Jan. 1989), pp. 1–13. DOI: 10.1101/sqb.1989.054.01.003. URL: https://doi.org/10.1101%2Fsqb.1989.054.01.003.

[89] CA Janeway, P Travers, and M Walport. "Antigen recognition by T cells". In: *Immunobiology: The Immune System in Health and Disease.* 5th ed. New York: Garland Science, 2001.

[90] Alice Jo et al. "The versatile functions of Sox9 in development, stem cells, and human diseases." In: *Genes & Diseases* 1.2 (Dec. 2014). Publisher: Elsevier, pp. 149–161. ISSN: 2352-4820. DOI: 10.1084/JEM.190.10.1427. URL: http://www.ncbi.nlm.nih.gov/pubmed/25685828.

[91] Arttu Jolma et al. "DNA-binding specificities of human transcription factors." eng. In: *Cell* 152.1-2 (Jan. 2013). Place: United States, pp. 327–339. ISSN: 1097-4172 0092-8674. DOI: 10.1016/j.cell.2012.12.009.

[92] Jonathan D. G. Jones and Jeffery L. Dangl. "The plant immune system". In: *Nature* 444.7117 (Nov. 2006), pp. 323–329. DOI: 10.1038/nature05286. URL: https://doi.org/10.1038%2Fnature05286.

[93] Adam D. Judge et al. "Interleukin 15 controls both proliferation and survival of a subset of memory-phenotype CD8(+) T cells." eng. In: *The Journal of experimental medicine* 196.7 (Oct. 2002), pp. 935–946. ISSN: 0022-1007 1540-9538. DOI: 10.1084/jem.20020772.

[94] Carl H June. "Adoptive T cell therapy for cancer in the clinic". eng. In: *The Journal of clinical investigation* 117.6 (June 2007). Publisher: American Society for Clinical Investigation, pp. 1466–1476. ISSN: 0021-9738. DOI: 10.1172/JCI32446. URL: https://pubmed.ncbi.nlm.nih.gov/17549249.

[95] jupyter-widgets. "ipywidgets". In: (2019). URL: https://github.com/jupyter-widgets/ipywidgets.

[96] Juhani Kähärä and Harri Lähdesmäki. "BinDNase: a discriminatory approach for transcription factor binding prediction using DNase I hypersensitivity data". In: *Bioinformatics* 31.17 (2015), pp. 2852–2859.

[97] James John Kaminski. "The Regulation of CD8+ T Cell Fate by Transcription Factor Binding and Chromatin Accessibility". PhD thesis. University of California, Berkeley, 2018.

[98] Donna Karolchik et al. "The UCSC Table Browser data retrieval tool." eng. In: *Nucleic Acids Res* 32.Database issue (Jan. 1, 2004), pp. D493–6. ISSN: 1362-4962 (Electronic); 0305-1048 (Print); 0305-1048 (Linking). DOI: 10.1093/nar/gkh103.

[99] A. Marijke Keestra et al. "Manipulation of small Rho GTPases is a pathogen-induced process detected by NOD1". In: *Nature* 496.7444 (Mar. 2013), pp. 233–237. DOI: 10.1038/nature12025. URL: https://doi.org/10.1038%2Fnature12025.

[100] Jens Keilwagen, Stefan Posch, and Jan Grau. "Accurate prediction of cell type-specific transcription factor binding". In: *Genome Biology* 20.1 (2019), p. 9.

[101] David R Kelley, Jasper Snoek, and John L Rinn. "Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks." eng. In: *Genome Res* 26.7 (July 2016), pp. 990–999. ISSN: 1549-5469 (Electronic); 1088-9051 (Print); 1088-9051 (Linking). DOI: 10.1101/gr.200535.115.

[102] Marion Kennedy et al. "T Lymphocyte Potential Marks the Emergence of Definitive Hematopoietic Progenitors in Human Pluripotent Stem Cell Differentiation Cultures". In: *Cell Reports* 2.6 (2012), pp. 1722–1735. ISSN: 2211-1247. DOI: https://doi.org/10.1016/j.celrep.2012.11.003. URL: https://www.sciencedirect.com/science/article/pii/S2211124712003841.

[103] HyeongJun Kim et al. "The Gene-Silencing Protein MORC-1 Topologically Entraps DNA and Forms Multimeric Assemblies to Cause DNA Compaction". In: *Molecular Cell* 75.4 (Aug. 2019), 700–710.e6. DOI: 10.1016/j.molcel.2019.07.032. URL: https://doi.org/10.1016%2Fj.molcel.2019.07.032.

[104] K Kim et al. "Epigenetic memory in induced pluripotent stem cells". eng. In: *Nature* 467.7313 (Sept. 2010), pp. 285–290. ISSN: 1476-4687. DOI: 10.1038/nature09342. URL: https://pubmed.ncbi.nlm.nih.gov/20644535.

[105] Kjetil Klepper et al. "Assessment of composite motif discovery methods". In: *BMC Bioinformatics* 9.1 (Feb. 2008), p. 123. ISSN: 1471-2105. DOI: 10.1186/1471-2105-9-123. URL: https://doi.org/10.1186/1471-2105-9-123.

[106] Thomas Kluyver et al. "Jupyter Notebooks: a publishing format for reproducible computational workflows". In: *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. Ed. by Fernando Loizides and Birgit Scmidt. IOS Press, 2016, pp. 87–90. URL: https://eprints.soton.ac.uk/403913/.

[107] Tony Kouzarides. "Chromatin modifications and their function." eng. In: *Cell* 128.4 (Feb. 23, 2007), pp. 693–705. ISSN: 0092-8674 (Print); 0092-8674 (Linking). DOI: 10.1016/j.cell.2007.02.005.

[108] A. Kundaje et al. "ENCODE-DREAM in vivo Transcription Factor Binding Site Prediction Challenge." In: *Synapse* (2017).

[109] Anshul Kundaje et al. "Integrative analysis of 111 reference human epigenomes". In: *Nature* 518.7539 (2015), pp. 317–330.

[110] Kornel Labun et al. "CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing". In: *Nucleic Acids Research* 47.W1 (May 2019), W171–W174. DOI: 10.1093/nar/gkz365. URL: https://doi.org/10.1093%2Fnar%2Fgkz365.

[111] Lindsay M LaFave et al. "Epigenomic State Transitions Characterize Tumor Progression in Mouse Lung Adenocarcinoma." eng. In: *Cancer Cell* 38.2 (Aug. 10, 2020), pp. 212–228. ISSN: 1878-3686 (Electronic); 1535-6108 (Print); 1535-6108 (Linking). DOI: 10.1016/j.ccell.2020.06.006.

[112] Ben Langmead and Steven L Salzberg. "Fast gapped-read alignment with Bowtie 2". In: *Nature Methods* 9.4 (2012), pp. 357–359. DOI: 10.1038/nmeth.1923. URL: https://doi.org/10.1038/nmeth.1923.

[113] Dongwon Lee. "LS-GKM: a new gkm-SVM for large-scale datasets". In: *Bioinformatics* 32.14 (2016), p. 2196. DOI: 10.1093/bioinformatics/btw142. eprint: /oup/backfile/Content_public/Journal/bioinformatics/32/14/10.1093_bioinformatics_btw142/2/btw142.pdf. URL: +%20http://dx.doi.org/10.1093/bioinformatics/btw142.

[114] Dongwon Lee et al. "A method to predict the impact of regulatory variants from DNA sequence". In: *Nature Genetics* 47.8 (2015), pp. 955–961.

[115] Tong Ihn Lee and Richard A. Young. "Transcription of eukaryotic protein-coding genes". In: *Annual Review of Genetics* 34.1 (2000). PMID: 11092823, pp. 77–137. DOI: 10.1146/annurev.genet.34.1.77. eprint: https://doi.org/10.1146/annurev.genet.34.1.77. URL: https://doi.org/10.1146/annurev.genet.34.1.77.

[116] Christina S Leslie, Eleazar Eskin, and William Stafford Noble. "The spectrum kernel: A string kernel for SVM protein classification." In: *Pacific Symposium on Biocomputing*. Vol. 7. 7. 2002, pp. 566–575.

[117] Bo Li and Colin N. Dewey. "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome". In: *BMC Bioinformatics* 12.1 (Aug. 2011), p. 323. ISSN: 1471-2105. DOI: 10.1186/1471-2105-12-323. URL: https://doi.org/10.1186/1471-2105-12-323.

[118] Sisi Li et al. "Mouse MORC3 is a GHKL ATPase that localizes to H3K4me3 marked chromatin". In: *Proceedings of the National Academy of Sciences* 113.35 (2016), E5108–E5116. ISSN: 0027-8424. DOI: 10.1073/pnas.1609709113. eprint: https://www.pnas.org/content/113/35/E5108.full.pdf. URL: https://www.pnas.org/content/113/35/E5108.

[119] Zhijian Li et al. "Identification of transcription factor binding sites using ATAC-seq." eng. In: *Genome Biol* 20.1 (Feb. 26, 2019), p. 45. ISSN: 1474-760X (Electronic); 1474-7596 (Print); 1474-7596 (Linking). DOI: 10.1186/s13059-019-1642-2.

[120] Yang Liao, Gordon K. Smyth, and Wei Shi. "featureCounts: an efficient general purpose program for assigning sequence reads to genomic features". In: *Bioinformatics* 30.7 (Nov. 2013), pp. 923–930. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btt656. eprint: https://academic.oup.com/bioinformatics/article-pdf/30/7/923/633148/btt656.pdf. URL: https://doi.org/10.1093/bioinformatics/btt656.

[121] Arthur Liberzon et al. "Molecular signatures database (MSigDB) 3.0." In: *Bioinformatics* 27.12 (May 2011). Publisher: Oxford University Press, pp. 1739–40. DOI: 10.1042/BJ2600463. URL: http://www.ncbi.nlm.nih.gov/pubmed/21546393.

[122] Giovanni Ligresti et al. "The acute phase reactant orosomucoid-1 is a bimodal regulator of angiogenesis with time- and context-dependent inhibitory and stimulatory properties." eng. In: *PloS one* 7.8 (2012), e41387. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0041387.

[123] Xiaoyu Liu et al. "Distinct features of H3K4me3 and H3K27me3 chromatin domains in pre-implantation embryos". In: *Nature* 537.7621 (2016), pp. 558–562. DOI: 10.1038/nature19362. URL: https://doi.org/10.1038/nature19362.

[124] Romain Lopez et al. "Deep generative modeling for single-cell transcriptomics". In: *Nature Methods* 15.12 (2018), pp. 1053–1058. DOI: 10.1038/s41592-018-0229-2. URL: https://doi.org/10.1038/s41592-018-0229-2.

[125] Zdravko J. Lorković et al. "Involvement of a GHKL ATPase in RNA-Directed DNA Methylation in Arabidopsis thaliana". In: *Current Biology* 22.10 (May 2012), pp. 933–938. DOI: 10.1016/j.cub.2012.03.061. URL: https://doi.org/10.1016%2Fj.cub.2012.03.061.

[126] Michael I. Love, Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome Biology* 15.12 (2014), p. 550. DOI: 10.1186/s13059-014-0550-8. URL: https://doi.org/10.1186/s13059-014-0550-8.

[127] Stefanie Luecke and Søren R. Paludan. "Innate Recognition of Alphaherpesvirus DNA". In: *Advances in Virus Research*. Elsevier, 2015, pp. 63–100. DOI: 10.1016/bs.aivir.2014.11.003. URL: https://doi.org/10.1016%2Fbs.aivir.2014.11.003.

[128] Hua Ma et al. "On use of partial area under the ROC curve for evaluation of diagnostic performance". In: *Statistics in medicine* 32.20 (2013), pp. 3449–3458. URL: https://pubmed.ncbi.nlm.nih.gov/23508757.

[129] Nimet Maherali et al. "A high-efficiency system for the generation and study of human induced pluripotent stem cells." eng. In: *Cell stem cell* 3.3 (Sept. 2008), pp. 340–345. ISSN: 1875-9777 1934-5909. DOI: 10.1016/j.stem.2008.08.003.

[130] Julien Mairal et al. "Convolutional kernel networks". In: *Advances in Neural Information Processing Systems*. 2014, pp. 2627–2635.

[131] Swapan Mallick et al. "The Simons Genome Diversity Project: 300 genomes from 142 diverse populations". In: *Nature* 538 (Sept. 2016). Article. URL: https://doi.org/10.1038/nature18964.

[132] "Mammalian Transcription Factor Networks: Recent Advances in Interrogating Biological Complexity". In: *Cell Systems* 5.4 (2017), pp. 319–331. URL: https://www.cell.com/cell-systems/fulltext/S2405-4712(17)30330-7?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS2405471217303307%3Fshowall%3Dtrue.

[133] A. Manning et al. "TOPMed Whole Genome Sequence Combined with Pancreas-Specific Annotation for Rare Variants Tests of Type 2 Diabetes Risk". In: *Diabetes* 67.Supplement 1 (2018). ISSN: 0012-1797. DOI: 10.2337/db18-1724-P. eprint: https://diabetes.diabetesjournals.org/content. URL: https://diabetes.diabetesjournals.org/content/67/Supplement_1/1724-P.

[134] Cory Y McLean et al. "GREAT improves functional interpretation of cis-regulatory regions". In: *Nature Biotechnology* 28.5 (May 2010), pp. 495–501. ISSN: 1546-1696. DOI: 10.1038/nbt.1630. URL: https://doi.org/10.1038/nbt.1630.

[135] Michael P. Meers, Dan Tenenbaum, and Steven Henikoff. "Peak calling by Sparse Enrichment Analysis for CUT&RUN chromatin profiling". In: *Epigenetics & Chromatin* 12.1 (2019), p. 42. DOI: 10.1186/s13072-019-0287-4. URL: https://doi.org/10.1186/s13072-019-0287-4.

[136] Wouter Meuleman et al. "Index and biological spectrum of human DNase I hypersensitive sites". In: *Nature* 584.7820 (2020), pp. 244–251.

[137] A. Chase Miller et al. "bam.iobio: a web-based, real-time, sequence alignment file inspector". In: *Nature Methods* 11 (Nov. 2014). Correspondence. URL: https://doi.org/10.1038/nmeth.3174.

[138] Yasuhiro Mimura et al. "Two-step colocalization of MORC3 with PML nuclear bodies". In: *Journal of Cell Science* 123.12 (June 2010), pp. 2014–2024. DOI: 10.1242/jcs.063586. URL: https://doi.org/10.1242%2Fjcs.063586.

[139] Liesbeth Minnoye et al. "Chromatin accessibility profiling methods". In: *Nature Reviews Methods Primers* 1.1 (2021), p. 10. DOI: 10.1038/s43586-020-00008-9. URL: https://doi.org/10.1038/s43586-020-00008-9.

[140] modin-project. "modin". In: (2019). URL: https://github.com/modin-project/modin.

[141] Jill E. Moore et al. "Expanded encyclopaedias of DNA elements in the human and mouse genomes". In: *Nature* 583.7818 (2020), pp. 699–710. DOI: 10.1038/s41586-020-2493-4. URL: https://doi.org/10.1038/s41586-020-2493-4.

[142] Alyssa Morrow. "Distributed Visualization for Genomic Analysis". MA thesis. EECS Department, University of California, Berkeley, May 2017. URL: http://www2.eecs.berkeley.edu/Pubs/TechRpts/2017/EECS-2017-82.html.

[143] Alyssa Morrow et al. "Convolutional Kitchen Sinks for Transcription Factor Binding Site Prediction". In: (May 2017).

[144] Alyssa Kramer Morrow et al. "Epitome: Predicting epigenetic events in novel cell types with multi-cell deep ensemble learning". In: *bioRxiv* (2021). DOI: 10.1101/2021.06.10.447140. eprint: https://www.biorxiv.org/content/early/2021/06/11/2021.06.10.447140.full.pdf. URL: https://www.biorxiv.org/content/early/2021/06/11/2021.06.10.447140.

[145] Alyssa Kramer Morrow et al. "Mango: Exploratory Data Analysis for Large-Scale Sequencing Datasets". In: *Cell Systems* 9.6 (2019), 609–613.e3. ISSN: 2405-4712. DOI: https://doi.org/10.1016/j.cels.2019.11.002. URL: https://www.sciencedirect.com/science/article/pii/S2405471219303886.

[146] Charlotte M Mousset et al. "Comprehensive Phenotyping of T Cells Using Flow Cytometry." In: *Cytometry Part A* 95.6 (Feb. 2019). Publisher: Wiley, pp. 647–654. DOI: 10.1002/CYTO.A.23724. URL: http://www.ncbi.nlm.nih.gov/pubmed/30714682.

[147] Charlotte M Mousset et al. "Ex vivo AKT-inhibition facilitates generation of polyfunctional stem cell memory-like CD8(+) T cells for adoptive immunotherapy". eng. In: *Oncoimmunology* 7.10 (Aug. 2018). Publisher: Taylor & Francis, e1488565–e1488565. ISSN: 2162-4011. DOI: 10.1080/2162402X.2018.1488565. URL: https://pubmed.ncbi.nlm.nih.gov/30288356.

[148] Armstrong Murira and Alain Lamarre. "Type-I Interferon Responses: From Friend to Foe in the Battle against Chronic Viral Infection". In: *Frontiers in Immunology* 7 (2016), p. 609. ISSN: 1664-3224. DOI: 10.3389/fimmu.2016.00609. URL: https://www.frontiersin.org/article/10.3389/fimmu.2016.00609.

[149] Seiji Nagano et al. "High Frequency Production of T Cell-Derived iPSC Clones Capable of Generating Potent Cytotoxic T Cells". In: *Molecular Therapy - Methods & Clinical Development* 16 (Mar. 2020). Publisher: Elsevier, pp. 126–135. ISSN: 2329-0501. DOI: 10.1016/j.omtm.2019.12.006. URL: https://doi.org/10.1016/j.omtm.2019.12.006 (visited on 07/15/2021).

[150] A S Narayanan, R C Page, and J. Swanson. "Collagen synthesis by human fibroblasts. Regulation by transforming growth factor-beta in the presence of other inflammatory mediators." In: *Biochemical Journal* 260.2 (June 1989). Publisher: Portland Press, pp. 463–9. ISSN: 0264-6021. DOI: 10.1093/BIOINFORMATICS/BTR260. URL: http://www.ncbi.nlm.nih.gov/pubmed/2504143.

[151] Sattva S. Neelapu et al. "Axicabtagene Ciloleucel CAR T-Cell Therapy in Refractory Large B-Cell Lymphoma." eng. In: *The New England journal of medicine* 377.26 (Dec. 2017), pp. 2531–2544. ISSN: 1533-4406 0028-4793. DOI: 10.1056/NEJMoa1707447.

[152] Shane Neph et al. "An expansive human regulatory lexicon encoded in transcription factor footprints". In: *Nature* 489.7414 (2012), pp. 83–90.

[153] Hong Hanh Nguyen et al. "Naïve CD8+ T cell derived tumor-specific cytotoxic effectors as a potential remedy for overcoming TGF-β immunosuppression in the tumor microenvironment". In: *Scientific Reports* 6.1 (June 2016), p. 28208. ISSN: 2045-2322. DOI: 10.1038/srep28208. URL: https://doi.org/10.1038/srep28208.

[154] Alexandros Nianias and Maria Themeli. "Induced Pluripotent Stem Cell (iPSC)-Derived Lymphocytes for Adoptive Cell Immunotherapy: Recent Advances and Challenges". eng. In: *Current hematologic malignancy reports* 14.4 (Aug. 2019). Publisher: Springer US, pp. 261–268. ISSN: 1558-822X. DOI: 10.1007/s11899-019-00528-6. URL: https://pubmed.ncbi.nlm.nih.gov/31243643.

[155] Hirofumi Noguchi, Chika Miyagi-Shiohira, and Yoshiki Nakashima. "Induced Tissue-Specific Stem Cells and Epigenetic Memory in Induced Pluripotent Stem Cells." eng. In: *International journal of molecular sciences* 19.4 (Mar. 2018). ISSN: 1422-0067. DOI: 10.3390/ijms19040930.

[156] Frank Austin Nothaft et al. "Rethinking Data-Intensive Science Using Scalable Analytics Systems". In: *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. SIGMOD '15. Melbourne, Victoria, Australia: ACM, 2015, pp. 631–646. DOI: 10.1145/2723372.2742787. URL: http://doi.acm.org/10.1145/2723372.2742787.

[157] Aisling O'Driscoll, Jurate Daugelaite, and Roy D. Sleator. "'Big data', Hadoop and cloud computing in genomics". In: *Journal of Biomedical Informatics* 46.5 (2013), pp. 774–781. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2013.07.001. URL: http://www.sciencedirect.com/science/article/pii/S1532046413001007.

[158] Shinya Oki et al. "ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data". In: *EMBO reports* 19.12 (2018), e46255. DOI: https://doi.org/10.15252/embr.201846255. eprint: https://www.embopress.org/doi/pdf/10.15252/embr.201846255. URL: https://www.embopress.org/doi/abs/10.15252/embr.201846255.

[159] Keisuke Okita, Tomoko Ichisaka, and Shinya Yamanaka. "Generation of germline-competent induced pluripotent stem cells". In: *Nature* 448.7151 (July 2007), pp. 313–317. ISSN: 1476-4687. DOI: 10.1038/nature05934. URL: https://doi.org/10.1038/nature05934.

[160] Mark Osborn et al. "CRISPR/Cas9 Targeted Gene Editing and Cellular Engineering in Fanconi Anemia." eng. In: *Stem cells and development* 25.20 (Oct. 2016), pp. 1591–1603. ISSN: 1557-8534 1547-3287. DOI: 10.1089/scd.2016.0149.

[161] Jae H. Park et al. "Long-Term Follow-up of CD19 CAR Therapy in Acute Lymphoblastic Leukemia." eng. In: *The New England journal of medicine* 378.5 (Feb. 2018), pp. 449–459. ISSN: 1533-4406 0028-4793. DOI: 10.1056/NEJMoa1709919.

[162] Georgios A Pavlopoulos et al. "Visualizing genome and systems biology: technologies, tools, implementation techniques and trends, past, present and future." eng. In: *Gigascience* 4 (2015), p. 38. ISSN: 2047-217X (Print); 2047-217X (Electronic); 2047-217X (Linking). DOI: 10.1186/s13742-015-0077-2.

[163] Jeffrey M. Perkel. *Challenge to scientists: does your ten-year-old code still run?* Ed. by Nature Technology Feature. [Online; posted 24-August-2020]. Aug. 2020. URL: %5Curl%7Bhttps://www.nature.com/articles/d41586-020-02462-7%7D.

[164] *Picard toolkit*. Publication Title: Broad Institute, GitHub repository. Broad Institute, 2018. URL: http://broadinstitute.github.io/picard/.

[165] Andreas Pichlmair and Caetano Reis e Sousa. "Innate recognition of viruses." eng. In: *Immunity* 27.3 (Sept. 26, 2007), pp. 370–383. ISSN: 1074-7613 (Print); 1074-7613 (Linking). DOI: 10.1016/j.immuni.2007.08.012.

[166] Bryan Quach and Terrence S Furey. "DeFCoM: analysis and modeling of transcription factor binding sites using a motif-centric genomic footprinter". In: *Bioinformatics* 33.7 (2016), pp. 956–963.

[167] Debasish Raha, Miyoung Hong, and Michael Snyder. "ChIP-Seq: A method for global identification of regulatory elements in the genome". In: *Current protocols in molecular biology* 91.1 (2010), pp. 21–19.

[168] Ali Rahimi and Benjamin Recht. "Random features for large-scale kernel machines". In: *Advances in Neural Information Processing Systems*. 2007, pp. 1177–1184.

[169] Ali Rahimi and Benjamin Recht. "Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning". In: *Advances in Neural Information Processing Systems*. 2009, pp. 1313–1320.

[170] Anil Raj et al. "msCentipede: Modeling Heterogeneity across Genomic Sites and Replicates Improves Accuracy in the Inference of Transcription Factor Binding". In: *PloS one* 10.9 (2015), e0138030–e0138030.

[171] Jeffrey Reid et al. *Genetic Variant-Phenotype Analysis System And Methods Of Use*. US Patent App. 15/473,302. Oct. 2017.

[172] Nicola Reynolds, Aoife O'Shaughnessy, and Brian Hendrich. "Transcriptional repressors: multifaceted regulators of gene expression." eng. In: *Development* 140.3 (Feb. 1, 2013), pp. 505–512. ISSN: 1477-9129 (Electronic); 0950-1991 (Linking). DOI: 10.1242/dev.083105.

[173] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." eng. In: *Bioinformatics* 26.1 (Jan. 1, 2010), pp. 139–140. ISSN: 1367-4811 (Electronic); 1367-4803 (Print); 1367-4803 (Linking). DOI: 10.1093/bioinformatics/btp616.

[174] T. James Robinson et al. "Integrative genomics viewer". In: *Nature biotechnology* 29 (2011), pp. 24–26. ISSN: 1087-0156. DOI: http://doi.org/10.1038/nbt.1754. URL: http://www.nature.com/nbt/journal/v29/n1/abs/nbt.1754.html.

[175] Anne Rogel et al. "Akt signaling is critical for memory CD8(+) T-cell development and tumor immune surveillance". eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 114.7 (Feb. 2017). Edition: 2017/01/30 Publisher: National Academy of Sciences, E1178–E1187. ISSN: 1091-6490. DOI: 10.1073/pnas.1611299114. URL: https://pubmed.ncbi.nlm.nih.gov/28137869.

[176] S. A. Rosenberg et al. "Observations on the systemic administration of autologous lymphokine-activated killer cells and recombinant interleukin-2 to patients with metastatic cancer." eng. In: *The New England journal of medicine* 313.23 (Dec. 1985). Place: United States, pp. 1485–1492. ISSN: 0028-4793. DOI: 10.1056/NEJM198512053132327.

[177] Michel Sadelain, Isabelle Rivière, and Stanley Riddell. "Therapeutic T cell engineering". eng. In: *Nature* 545.7655 (May 2017), pp. 423–431. ISSN: 1476-4687. DOI: 10.1038/nature22395. URL: https://pubmed.ncbi.nlm.nih.gov/28541315.

[178] Neville E Sanjana, Ophir Shalem, and Feng Zhang. "Improved vectors and genome-wide libraries for CRISPR screening". In: *Nature Methods* 11.8 (July 2014), pp. 783–784. DOI: 10.1038/nmeth.3047. URL: https://doi.org/10.1038%2Fnmeth.3047.

[179] Ansuman T Satpathy et al. "Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion." eng. In: *Nat Biotechnol* 37.8 (Aug. 4, 2019), pp. 925–936. ISSN: 1546-1696 (Electronic); 1087-0156 (Print); 1087-0156 (Linking). DOI: 10.1038/s41587-019-0206-z.

[180] Michael C. Schatz. "CloudBurst: highly sensitive read mapping with MapReduce". In: *Bioinformatics* (June 2009).

[181] Alicia N Schep et al. "chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data". In: *Nature Methods* 14.10 (2017), pp. 975–978. DOI: 10.1038/nmeth.4401. URL: https://doi.org/10.1038/nmeth.4401.

[182] Tobias Schmidt, Jonathan L. Schmid-Burgk, and Veit Hornung. "Synthesis of an arrayed sgRNA library targeting the human genome". In: *Scientific Reports* 5.1 (Oct. 2015). DOI: 10.1038/srep14987. URL: https://doi.org/10.1038%2Fsrep14987.

[183] Tobias Schmidt et al. "Designer Nuclease-Mediated Generation of Knockout THP1 Cells". In: *TALENs*. Springer New York, 2016, pp. 261–272. DOI: 10.1007/978-1-4939-2932-0_19. URL: https://doi.org/10.1007%2F978-1-4939-2932-0_19.

[184] S. Schneeweiss. "Improving therapeutic effectiveness and safety through big healthcare data". In: *Clinical Pharmacology* 99 (3 2016). URL: https://ascpt.onlinelibrary.wiley.com/doi/abs/10.1002/cpt.316.

[185] Jacob Schreiber et al. "A pitfall for machine learning methods aiming to predict across cell types". In: *bioRxiv* (2019). eprint: https://www.biorxiv.org/content/early/2019/01/04/512434.full.pdf. URL: https://www.biorxiv.org/content/early/2019/01/04/512434.

[186] Jacob Schreiber et al. "Avocado: a multi-scale deep tensor factorization method learns a latent representation of the human epigenome". In: *Genome Biology* 21.1 (2020), p. 81.

[187] Roy Schwartz et al. "Green AI". In: *CoRR* abs/1907.10597 (2019). arXiv: 1907.10597. URL: http://arxiv.org/abs/1907.10597.

[188] Debattama R. Sen et al. "The epigenetic landscape of T cell exhaustion". In: *Science* 354.6316 (2016). Publisher: American Association for the Advancement of Science _eprint: https://science.sciencemag.org/content/354/6316/1165.full.pdf, pp. 1165–1169. ISSN: 0036-8075. DOI: 10.1126/science.aae0491. URL: https://science.sciencemag.org/content/354/6316/1165.

[189] "SETDB1: a novel KAP-1-associated histone H3, lysine 9-specific methyltransferase that contributes to HP1-mediated silencing of euchromatic genes by KRAB zinc-finger proteins". In: *Genes & development* (2002). DOI: 10.1101/gad.973302. URL: https://pubmed.ncbi.nlm.nih.gov/11959841/.

[190] M. Setty and CS Leslie. "SeqGL identifies context-dependent binding signals in genome-wide regulatory element maps". In: *PLoS Computational Biology* 11.5 (2015), e1004271.

[191] Jay Shendure and Hanlee Ji. "Next-generation DNA sequencing". In: *Nature Biotechnology* 26 (Oct. 2008). URL: https://doi.org/10.1038/nbt1486.

[192] Wenjie Shu et al. "Genome-wide analysis of the relationships between DNaseI HS, histone modifications and gene expression reveals distinct modes of chromatin domains". In: *Nucleic acids research* 39.17 (2011), pp. 7428–7443.

[193] Peter J Skene and Steven Henikoff. "An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites". In: *eLife* 6 (Jan. 2017). Ed. by Danny Reinberg, e21856. ISSN: 2050-084X.

[194] E. Mitchell Skinner et al. "JBrowse: A next-generation genome browser". In: *Genome Research* 19 (2009).

[195] Elizabeth Sloan et al. "Analysis of the SUMO2 Proteome during HSV-1 Infection". In: *PLOS Pathogens* 11.7 (July 2015). Ed. by Paul D. Ling, e1005059. DOI: 10.1371/journal.ppat.1005059. URL: https://doi.org/10.1371%2Fjournal.ppat.1005059.

[196] M J Solomon and A Varshavsky. "Formaldehyde-mediated DNA-protein crosslinking: a probe for in vivo chromatin structures." eng. In: *Proc Natl Acad Sci U S A* 82.19 (Oct. 1, 1985), pp. 6470–6474. ISSN: 0027-8424 (Print); 1091-6490 (Electronic); 0027-8424 (Linking). DOI: 10.1073/pnas.82.19.6470.

[197] C L Sommers et al. "A role for the Tec family tyrosine kinase Txk in T cell activation and thymocyte selection." In: *The Journal of Experimental Medicine* 190.10 (Nov. 1999). Publisher: The Rockefeller University Press, pp. 1427–38. ISSN: 0022-1007. DOI: 10.1016/J.IMMUNI.2008.09.019. URL: http://www.ncbi.nlm.nih.gov/pubmed/10562318.

[198] Philipp N. Spahn et al. "PinAPL-Py: A comprehensive web-application for the analysis of CRISPR/Cas9 screens". In: *Scientific Reports* 7.1 (Nov. 2017). DOI: 10.1038/s41598-017-16193-9. URL: https://doi.org/10.1038%2Fs41598-017-16193-9.

[199] Judith Staerk et al. "Reprogramming of human peripheral blood cells to induced pluripotent stem cells." eng. In: *Cell stem cell* 7.1 (July 2010), pp. 20–24. ISSN: 1875-9777 1934-5909. DOI: 10.1016/j.stem.2010.06.002.

[200] Trevor Standley et al. "Which Tasks Should Be Learned Together in Multi-task Learning?" In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 9120–9132.

[201] Spyridon Stavrou and Susan R Ross. "APOBEC3 Proteins in Viral Immunity." In: *The Journal of Immunology* 195.10 (Nov. 2015). Publisher: The American Association of Immunologists, pp. 4565–70. DOI: 10.1016/J.GENDIS.2014.09.004. URL: http://www.ncbi.nlm.nih.gov/pubmed/26546688.

[202] Emma Strubell, Ananya Ganesh, and Andrew McCallum. "Energy and Policy Considerations for Deep Learning in NLP". In: *CoRR* abs/1906.02243 (2019). arXiv: 1906.02243. URL: http://arxiv.org/abs/1906.02243.

[203] Kazutoshi Takahashi and Shinya Yamanaka. "Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors". In: *Cell* 126.4 (Aug. 2006), pp. 663–676. ISSN: 0092-8674. DOI: 10.1016/j.cell.2006.07.024. URL: https://www.sciencedirect.com/science/article/pii/S0092867406009767.

[204] Keiko Takahashi et al. "Dynamic Regulation of p53 Subnuclear Localization and Senescence by MORC3". In: *Molecular Biology of the Cell* 18.5 (May 2007). Ed. by A. Gregory Matera, pp. 1701–1709. DOI: 10.1091/mbc.e06-08-0747. URL: https://doi.org/10.1091%2Fmbc.e06-08-0747.

[205] Mingxing Tan and Quoc V. Le. "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks". In: *CoRR* abs/1905.11946 (2019). arXiv: 1905.11946. URL: http://arxiv.org/abs/1905.11946.

[206] Rohan Taori et al. *Measuring Robustness to Natural Distribution Shifts in Image Classification*. 2020. arXiv: 2007.00644 [cs.LG].

[207] Iva A Tchasovnikarova et al. "Hyperactivation of HUSH complex function by Charcot–Marie–Tooth disease mutation in MORC2". In: *Nature Genetics* 49.7 (June 2017), pp. 1035–1044. DOI: 10.1038/ng.3878. URL: https://doi.org/10.1038%2Fng.3878.

[208] Leonid Teytelman et al. "Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins". In: *Proceedings of the National Academy of Sciences* 110.46 (2013), pp. 18602–18607. ISSN: 0027-8424. DOI: 10.1073/pnas.1316064110. eprint: https://www.pnas.org/content/110/46/18602.full.pdf. URL: https://www.pnas.org/content/110/46/18602.

[209] The Cancer Genome Atlas Research Network et al. "The Cancer Genome Atlas Pan-Cancer analysis project". In: *Nature Genetics* 45 (Sept. 2013), pp. 1113–1120. URL: https://doi.org/10.1038/ng.2764.

[210] Maria Themeli, Isabelle Rivière, and Michel Sadelain. "New cell sources for T cell engineering and adoptive immunotherapy." eng. In: *Cell stem cell* 16.4 (Apr. 2015), pp. 357–366. ISSN: 1875-9777 1934-5909. DOI: 10.1016/j.stem.2015.03.011.

[211] Peter J Thompson, Todd S Macfarlan, and Matthew C Lorincz. "Long Terminal Repeats: From Parasitic Elements to Building Blocks of the Transcriptional Regulatory Repertoire." eng. In: *Mol Cell* 62.5 (June 2016), pp. 766–776. ISSN: 1097-4164 (Electronic); 1097-2765 (Print); 1097-2765 (Linking). DOI: 10.1016/j.molcel.2016.03.029.

[212] Robert E. Thurman et al. "The accessible chromatin landscape of the human genome". In: *Nature* 489.7414 (2012), pp. 75–82. DOI: 10.1038/nature11232. URL: https://doi.org/10.1038/nature11232.

[213] Jakub Tolar et al. "Induced pluripotent stem cells from individuals with recessive dystrophic epidermolysis bullosa." eng. In: *The Journal of investigative dermatology* 131.4 (Apr. 2011), pp. 848–856. ISSN: 1523-1747 0022-202X. DOI: 10.1038/jid.2010.346.

[214] Jakub Tolar et al. "Keratinocytes from induced pluripotent stem cells in junctional epidermolysis bullosa." eng. In: *The Journal of investigative dermatology* 133.2 (Feb. 2013), pp. 562–565. ISSN: 1523-1747 0022-202X. DOI: 10.1038/jid.2012.278.

[215] Jakub Tolar et al. "Patient-specific naturally gene-reverted induced pluripotent stem cells in recessive dystrophic epidermolysis bullosa". eng. In: *The Journal of investigative dermatology* 134.5 (May 2014). Edition: 2013/12/06, pp. 1246–1254. ISSN: 1523-1747. DOI: 10.1038/jid.2013.523. URL: https://pubmed.ncbi.nlm.nih.gov/24317394.

[216] "Transcription Factors". In: *Reference Module in Biomedical Sciences*. Elsevier, 2014. DOI: https://doi.org/10.1016/B978-0-12-801238-3.05466-0. URL: https://www.sciencedirect.com/science/article/pii/B9780128012383054660.

[217] "Transcription factors and evolution: An integral part of gene expression (Review)". In: *World Academy of Sciences Journal* 2.1 (2020), pp. 3–8. URL: https://www.spandidos-publications.com/10.3892/wasj.2020.32.

[218] Dan Vanderkam et al. "pileup.js: a JavaScript library for interactive and in-browser visualization of genomic data". In: *Bioinformatics* 32.15 (Aug. 2016). ISSN: 13674803. DOI: 10.1093/bioinformatics/btw167. URL: http://dx.doi.org/10.1093/bioinformatics/btw167.

[219] J. and Vaquerizas. "A census of human transcription factors: function, expression and evolution". In: *Nature Reviews Genetics* (2009), pp. 252–263. DOI: https://doi.org/10.1038/nrg2538. URL: https://www.nature.com/articles/nrg2538.

[220] Jeff Vierstra and John A Stamatoyannopoulos. "Genomic footprinting". In: *Nature Methods* 13.3 (2016), pp. 213–221.

[221] Pauli Virtanen et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* (2020). DOI: https://doi.org/10.1038/s41592-019-0686-2.

[222] Raul Vizcardo et al. "Regeneration of Human Tumor Antigen-Specific T Cells from iPSCs Derived from Mature CD8+ T Cells". In: *Cell Stem Cell* 12.1 (Jan. 2013). Publisher: Elsevier, pp. 31–36. ISSN: 1934-5909. DOI: 10.1016/j.stem.2012.12.006. URL: https://doi.org/10.1016/j.stem.2012.12.006 (visited on 07/15/2021).

[223] Kai Wähner. *Data Preprocessing vs. Data Wrangling in Machine Learning Projects*. Mar. 2017. URL: https://www.infoq.com/articles/ml-data-processing/.

[224] Lie Wang et al. "The zinc finger transcription factor Zbtb7b represses CD8-lineage gene expression in peripheral CD4+ T cells." In: *Immunity* 29.6 (Dec. 2008). Publisher: Elsevier, pp. 876–87. DOI: 10.4049/JIMMUNOL.1501504. URL: http://www.ncbi.nlm.nih.gov/pubmed/19062319.

[225] Matthew T. Weirauch et al. "Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity". In: *Cell* 158.6 (2014), pp. 1431–1443. ISSN: 0092-8674. DOI: https://doi.org/10.1016/j.cell.2014.08.009. URL: https://www.sciencedirect.com/science/article/pii/S0092867414010368.

[226] Felix M Wensveen, Vedrana Jelencic, and Bojan Polic. "NKG2D: A Master Regulator of Immune Cell Responsiveness." In: *Frontiers in Immunology* 9 (Jan. 2018). Publisher: Frontiers Media S.A. ISSN: 1664-3224. DOI: 10.3389/FIMMU.2018.00441. URL: http://www.ncbi.nlm.nih.gov/pubmed/29568297.

[227] Anna-Lena Van de Weyer et al. "A Species-Wide Inventory of NLR Genes and Alleles in Arabidopsis thaliana". In: *Cell* 178.5 (Aug. 2019), 1260–1272.e14. DOI: 10.1016/j.cell.2019.07.038. URL: https://doi.org/10.1016%2Fj.cell.2019.07.038.

[228] *What is the epigenome?* URL: https://www.futurelearn.com/info/courses/the-genomics-era/0/steps/4875.

[229] Warren A Whyte et al. "Enhancer decommissioning by LSD1 during embryonic stem cell differentiation." eng. In: *Nature* 482.7384 (Feb. 1, 2012), pp. 221–225. ISSN: 1476-4687 (Electronic); 0028-0836 (Print); 0028-0836 (Linking). DOI: 10.1038/nature10805.

[230] Matthew A Williams and Michael J Bevan. "Effector and memory CTL differentiation." In: *Annual Review of Immunology* 25 (Jan. 2007). Publisher: Annual Reviews, pp. 171–92. ISSN: 0732-0582. DOI: 10.1146/ANNUREV.IMMUNOL.25.022106.141548. URL: http://www.ncbi.nlm.nih.gov/pubmed/17129182.

[231] Hao Xu et al. "Innate immune sensing of bacterial modifications of Rho GTPases by the Pyrin inflammasome". In: *Nature* 513.7517 (June 2014), pp. 237–241. DOI: 10.1038/nature13449. URL: https://doi.org/10.1038%2Fnature13449.

[232] Feng Yan et al. "From reads to insight: a hitchhiker's guide to ATAC-seq data analysis". In: *Genome Biology* 21.1 (2020), p. 22. DOI: 10.1186/s13059-020-1929-3. URL: https://doi.org/10.1186/s13059-020-1929-3.

[233] Qijin Yin et al. "DeepHistone: a deep learning approach to predicting histone modifications." In: *BMC Genomics* 20.Suppl 2 (Apr. 4, 2019), p. 193.

[234] Guangchuang Yu et al. "clusterProfiler: an R package for comparing biological themes among gene clusters." eng. In: *Omics : a journal of integrative biology* 16.5 (May 2012), pp. 284–287. ISSN: 1557-8100 1536-2310. DOI: 10.1089/omi.2011.0118.

[235] Han Yuan et al. "BindSpace decodes transcription factor binding signals by large-scale sequence embedding". In: *Nature Methods* 16.9 (2019), pp. 858–861.

[236] Matei Zaharia et al. "Resilient Distributed Datasets: A Fault-tolerant Abstraction for In-memory Cluster Computing". In: *Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation*. NSDI'12. San Jose, CA: USENIX Association, 2012, pp. 15–28. URL: http://dl.acm.org/citation.cfm?id=2228298.2228301.

[237] Hao Zhang, Pu Zhao, and He Huang. "Engineering better chimeric antigen receptor T cells". In: *Experimental Hematology & Oncology* 9.1 (Dec. 2020), p. 34. ISSN: 2162-3619. DOI: 10.1186/s40164-020-00190-2. URL: https://doi.org/10.1186/s40164-020-00190-2.

[238] Hong Zhang et al. "The NEI/NCBI dbGAP database: genotypes and haplotypes that may specifically predispose to risk of neovascular age-related macular degeneration." In: *BMC medical genetics* 9 (June 2008), p. 51. ISSN: 1471-2350. DOI: 10.1186/1471-2350-9-51. URL: http://dx.doi.org/10.1186/1471-2350-9-51.

[239] Yong Zhang et al. "Model-based Analysis of ChIP-Seq (MACS)". In: *Genome Biology* 9.9 (2008), R137. DOI: 10.1186/gb-2008-9-9-r137. URL: https://doi.org/10.1186/gb-2008-9-9-r137.

[240] Jian Zhou and Olga G Troyanskaya. "Predicting effects of noncoding variants with deep learning-based sequence model". In: *Nature methods* 12.10 (2015), pp. 931–934.

[241] Qian Zhu et al. "CUT&RUNTools: a flexible pipeline for CUT&RUN processing and footprint analysis". In: *Genome Biology* 20.1 (2019), p. 192. DOI: 10.1186/s13059-019-1802-4. URL: https://doi.org/10.1186/s13059-019-1802-4.