Title
The Regulation of CD8+ T Cell Fate by Transcription Factor Binding and Chromatin Accessibility

Author
Kaminski, James

Publication Date
2018

Peer reviewed|Thesis/dissertation

The Regulation of CD8+ T Cell Fate by Transcription Factor Binding and Chromatin Accessibility

By

James John Kaminski

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Computational Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Nir Yosef, Chair
Professor Ellen Robey
Professor Lexin Li
Professor Donald C. Rio

Summer 2018

Abstract

The Regulation of CD8+ T Cell Fate by Transcription Factor Binding and Chromatin Accessibility

by

James John Kaminski

Doctor of Philosophy in Computational Biology

University of California, Berkeley

Professor Nir Yosef, Chair

CD8+ T cells fight viral infections and other intracellular pathogens by identifying and destroying infected cells. CD8+ T cells, like all cells, take on different cell states which determine their behavior. This dissertation examines how epigenetic mechanisms such as chromatin accessibility and transcription factor binding regulate transcription, and consequently CD8+ T cell state. The particular CD8+ T cell states that I examine in this dissertation are: Naive (quiescent cells), Effector (cells actively fighting infection), Memory (cells that remain after infection is cleared, ready to fight recurrence of same infection), and Exhausted (cells with weakened cytotoxic function due to prolonged encounter with a particular antigen).

Chapter One introduces CD8+ T cells, regulation of transcription, and some of the high-throughput sequencing assays used later in the text to study CD8+ T cells. Chapter Two describes some of the computational tools I developed to process and infer useful information from high-throughput sequencing data. Chapter Three presents evidence that the chromatin of Exhausted CD8+ T cells differs from that of normal T cells, and demonstrates that these differences have functional consequences by validating an exhaustion-specific enhancer for PD-1, an important immunotherapy target. Chapter Four presents evidence that Batf and Irf4 work together to remodel chromatin and affect the binding of key CD8+ T cell transcription factors during the transition from naive to effector cell state.

To Molly

Acknowledgements

I am a lucky person. I have received a tremendous amount of help throughout this dissertation, and throughout my life. Below, I briefly acknowledge some of the many people who have helped me, but a full accounting would be longer than the dissertation itself. My apologies to those I did not mention here.

To Molly, my parents, and siblings, thank you for your many years of support, patient listening, and encouragement.

To Nir, and my colleagues in the Yosef Lab, thank you for all your guidance, for helping me when I struggled, for bringing me onto exciting research projects, and for being a constant source of inspiration.

To the Huttenhower Lab and the Urban Institute, thank you for giving me my start in research, teaching me the basics, and for being patient with me as I learned.

To Deb Sen, Hsiao-Wei Tsao, Nick Haining and the rest of the Haining Lab, thank you for including me on these wonderful research projects, and for setting an example through your tremendous work ethic and enthusiasm for research.

To the computational biology program at Berkeley, thank you for taking a chance and accepting me, and thank you to the kind and generous faculty, administrative staff, and fellow students for all your help.

To my friends at Berkeley, in DC, and elsewhere, for all your kind words, constant support, and hilarious jokes.

To my dissertation committee, thank you for your encouragement, for reading my drafts, offering advice, and for taking time out of your busy schedules to guide and advise me.

To the NIH and the people who pay taxes to fund it, thank you for supporting my research. I hope that our improved understanding of T cells will lead to better treatment for those suffering from cancer and other diseases.

Chapter 1 - Introduction

1.1 CD8+ T cells

The immune system employs a remarkable array of mechanisms to protect an organism from infection. Mechanisms which are highly specific are classified as belonging to the adaptive immune system, and those which are more general belong to the innate immune system. B cells, for example, produce antibodies that bind tightly to a particular pathogen or its product, and much less effectively to other substances, and so belong to the adaptive immune system.

CD8+ T cells are part of the adaptive immune system and target viral infections, other intracellular pathogens, and cancerous cells. T cells fend off these pathogenic cells by secreting proteins which interfere with viral replication such as interferon-gamma (Murphy 2011) and by inducing targeted cells to undergo apoptosis. The highly specific nature of a T cell derives from the T cell receptor (TCR), which recognizes particular antigens - substances that trigger an immune response (a viral protein, for example) - by binding tightly to one type of antigen and weakly to other molecules. Infected cells enable this recognition by posting portions of viral proteins on their surface in MHC Class I molecules. CD8+ T cells with TCR's specific to that antigen will bind tightly to this combination of MHC Class I molecule and viral protein, secrete perforin to create a hole in the membrane of the infected cell, and then release granzymes which direct the targeted cell to undergo apoptosis.

CD8+ T cells, like all cells, take on different cell states which determine their behavior. For example, a "naive" CD8+ T cell (which has not been primed by an antigen-presenting cell for effector function) will fail to destroy an infected cell upon encounter. **The central aim of this dissertation is to understand how epigenetic mechanisms such as chromatin accessibility and transcription factor binding regulate transcription, and consequently CD8+ T cell state.** The particular CD8+ T cell states that I examine in this dissertation are: Naive (quiescent cells), Effector (cells actively fighting infection), Memory (cells that remain after infection is cleared, ready to fight recurrence of same infection), and Exhausted (cells with weakened cytotoxic function due to prolonged encounter with a particular antigen). Cells in each of these states feature particular patterns of gene expression which direct their function. I briefly review these states and their functional consequences below.

Lymphoid progenitor cells originate in the bone marrow and then proceed to the thymus where they can develop into T cells (among other possible fates). At the end of this process, CD8+ T cells with a functional CD8 protein and TCR will then migrate throughout the body. At this stage, the cells are **Naive**, and must receive additional signals to engage their cytotoxic capabilities. A dendritic cell or other antigen-presenting cell activates T cells by presenting the cognate antigen to the T cell along with co-stimulatory signals. Activated T cells are **Effector** T cells, which divide rapidly, secrete

the cytokines TNF-alpha and IFN-gamma (N. Zhang and Bevan 2011) **,** and destroy infected cells with perforin and granzymes. After an infection is cleared, a subpopulation of T cells will remain behind as **Memory** cells, which can quickly combat a recurrence of their target pathogen. In cases where the infection is not cleared, CD8+ T cells can become **Exhausted,** display weakened capability to combat infection, and will themselves undergo apoptosis.

1.2 Transcription Factors Bind to Enhancers to Regulate Gene Expression and Cell Fate

Each of the T cell states described above exhibit different functions despite being the same type of cell. Thinking about this phenomenon in a wider context, we know that (with some small exceptions), every cell in an organism shares an identical genetic code, yet this same collection of cells varies dramatically in appearance, function, and behavior. How is this possible? How can a rod cell in the retina selectively produce rhodopsin, a protein for sensing different wavelengths of light, while an effector CD8+ T cell selectively produces perforin for boring a hole in an infected cell?

The answer is that individual cells have mechanisms for selectively choosing which genes are transcribed and the rate at which they are transcribed - a set of mechanisms we call "regulation of transcription". In order to understand how T cells alter their transcriptional program to switch from one fate to another, I studied two particular regulatory mechanisms: 1) the opening and closing of chromatin (DNA and the proteins attached it), and 2) the binding of transcription factors to enhancers.

DNA is quite long relative to the dimensions of eukaryotic cells (two meters in length, compared to the roughly 6 to 7 um diameter of a typical human lymphocyte[1]), and is able to fit inside the nucleus only by intricate folding and packing. The lowest level of organization is the nucleosome - where 147bp on DNA are tightly wrapped around octamers of histone proteins. The nucleosomes themselves can be tightly packed together and condensed. The importance of this with respect to transcription is that DNA which is tightly packaged into dense units of nucleosomes (heterochromatin) is less accessible to the RNA Pol II machinery, and consequently will not be transcribed or will only be transcribed at low levels. In contrast, stretches of DNA with few nucleosomes (euchromatin) are more open and easier for the RNA Polymerase machinery to bind to and transcribe. Proteins called chromatin remodelers can bind to DNA, modify the attachment of histone tails to DNA, and consequently open/close regions of DNA. We are able to measure the accessibility of particular regions of DNA with the ATAC-Seq protocol (Buenrostro et al. 2013)**,** which I discuss in more detail later.

[1] This value was calculated assuming a roughly spherical lymphocyte with a volume of 130 um^3. This number was obtained was from the Bionumbers Database (Milo et al. 2010) [BNID: 111439], which rounded an estimate of from (Schmid-Schönbein, Shih, and Chien 1980).

The RNA polymerase may encounter regions of DNA by chance and transcribe them at a low or "basal" level, but additional mechanisms are required to transcribe mRNA at the degree necessary for cells to carry out their state-specific functions. Proteins called transcription factors (TFs) bind to the DNA and carry out numerous regulatory functions such as recruiting Pol II to a promoter of a particular gene, recruiting other factors which recruit Pol II, recruiting the chromatin remodeling machinery, linking the promoter of a key gene to an enhancer far away on the genome, and other actions to increase/decrease transcription of a particular gene. We identify where proteins are bound on the genome using the ChIP-Seq protocol (Johnson et al. 2007), and also through TF footprinting of ATAC-Seq data. (The latter method attempts to identify where a TF is bound based on the particular pattern of the protocol's Tn5 transposase binding and cutting around it.)

In addition to chromatin accessibility and TF binding, many other mechanisms regulate the distribution of mRNA and proteins produced by the cell. Pol II pausing for example, can also affect the rate of transcription, and post-transcriptional mechanisms such as alternative splicing affect the final distribution of proteins produced by a cell in a given state. These are outside the scope of this dissertation, but are important to keep in mind.

1.3 High-throughput Sequencing Methods Are Well-Suited for Studying Regulation of Transcription

To identify possible enhancers, we search for regions of DNA that are accessible in a particular cell type. For this, we use ATAC-Seq, which uses a Tn5 transposase to fragment the genome and prepare it for sequencing. Areas of the genome which are more accessible will be cut by the transposase at higher rates, and consequently there will be more fragments mapping to those regions in the sequencing data. Users can align the data to the reference genome for their organism, and then use "peak calling" methods such as MACS2 (Y. Zhang et al. 2008) to identify regions with particularly high pileups of fragments.

The ChIP-Seq protocol is used to identify where particular transcription factors are bound to the genome. The protocol consists of taking a large population of cells, treating them with formaldehyde so all of their proteins stay bound to the DNA, fragmenting the DNA with a method such as sonication, and then using a probe to retrieve fragments bound with the protein of interest. (This can be done with antibodies conjugated to a magnetic bead, for example.) As with ATAC-Seq, users can map these reads to the genome and call peaks to identify regions likely bound by their protein of interest. Researchers typically compare their data with a control sample with no antibody treatment, to exclude regions that tend to turn up in sequencing experiments due to bias.

Once we know where the enhancers are, and what proteins are bound to them, we are curious to know which genes they are regulating. One might be tempted to assume they are regulating the nearest gene, but as recent reviews of enhancers note (Vernimmen D 2015) (Levine, Cattoglio, and Tjian 2014), there are cases of enhancers regulating a gene over a million base pairs away. Using chromatin conformation methods like HiChIP (Mumbach et al. 2016), we can identify long-range contacts on the genome and interactions between enhancers and promoters of associated genes. Also, we can use CRISPR-Cas9 (Jinek et al. 2012) to mutate and disable enhancers (Canver et al. 2015) and then measure expression of their putative target genes.

1.4 What are computational biologists good for?

The contributions of the "wetlab" scientists are abundantly clear - without them, there would be no data, and they are the experts in molecular biology. Their careful hands are the ones that actually carry out all of the high-throughput experiments described above. [I have been fortunate to take a lab course at Berkeley, and also assist once with retrieving T cells, sorting them, and extracting RNA - so I have some sense of the tremendous skill and work ethic possessed by my colleagues.]

A reader of this dissertation may fairly ask, "Well… what do they need you for?", and I can hopefully justify myself. Computational biologists have training in statistics, and we use this knowledge to assist in the planning of experiments so there is sufficient power to test a particular hypothesis, identify the appropriate method / statistical test for research questions that emerge during a project, and develop new methods when they are needed. Computational biologists also have training in computer science, and develop new algorithms to efficiently process large amounts of sequencing data in parallel, automate time-consuming tasks, and infer useful information from heterogeneous and complex datasets from multiple experiments. And finally, and most importantly, we also have training in biology, and use that knowledge to conceive and test hypotheses for our research projects and make real scientific contributions alongside our collaborators. My training at Berkeley has prepared me well in all three areas, and my teaching and collaborative experience taught me how to clearly explain ideas in one field to practitioners in one of the other two. In the next chapter, I present some of the computational methods I have developed, and discuss their application to scientific questions in the subsequent chapters.

1.5 References

Buenrostro, Jason D., Paul G. Giresi, Lisa C. Zaba, Howard Y. Chang, and William J. Greenleaf. 2013. "Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-Binding Proteins and Nucleosome Position." *Nature Methods* 10 (12): 1213–18.

Canver, Matthew C., Elenoe C. Smith, Falak Sher, Luca Pinello, Neville E. Sanjana, Ophir Shalem, Diane D. Chen, et al. 2015. "BCL11A Enhancer Dissection by Cas9-Mediated in

Situ Saturating Mutagenesis." *Nature* 527 (7577): 192–97.
Jinek, Martin, Krzysztof Chylinski, Ines Fonfara, Michael Hauer, Jennifer A. Doudna, and Emmanuelle Charpentier. 2012. "A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity." *Science* 337 (6096): 816–21.
Johnson, David S., Ali Mortazavi, Richard M. Myers, and Barbara Wold. 2007. "Genome-Wide Mapping of in Vivo Protein-DNA Interactions." *Science* 316 (5830): 1497–1502.
Levine, Michael, Claudia Cattoglio, and Robert Tjian. 2014. "Looping back to Leap Forward: Transcription Enters a New Era." *Cell* 157 (1): 13–25.
Milo, Ron, Paul Jorgensen, Uri Moran, Griffin Weber, and Michael Springer. 2010. "BioNumbers--the Database of Key Numbers in Molecular and Cell Biology." *Nucleic Acids Research* 38 (Database issue): D750–53.
Mumbach, Maxwell R., Adam J. Rubin, Ryan A. Flynn, Chao Dai, Paul A. Khavari, William J. Greenleaf, and Howard Y. Chang. 2016. "HiChIP: Efficient and Sensitive Analysis of Protein-Directed Genome Architecture." *Nature Methods* 13 (11): 919–22.
Murphy, Kenneth. 2011. *Janeway's Immunobiology by Murphy,Kenneth*. Garland Sceince,2011. 8th Edition.
Schmid-Schönbein, G. W., Y. Y. Shih, and S. Chien. 1980. "Morphometry of Human Leukocytes." *Blood* 56 (5): 866–75.
Vernimmen D. 2015. "The Hierarchy of Transcriptional Activation: From Enhancer to Promoter." *Trends in Genetics: TIG*, December. http://www.ncbi.nlm.nih.gov/pubmed/26599498.
Zhang, Nu, and Michael J. Bevan. 2011. "CD8+ T Cells: Foot Soldiers of the Immune System." *Immunity* 35 (2): 161–68.
Zhang, Yong, Tao Liu, Clifford A. Meyer, Jérôme Eeckhoute, David S. Johnson, Bradley E. Bernstein, Chad Nusbaum, et al. 2008. "Model-Based Analysis of ChIP-Seq (MACS)." *Genome Biology* 9 (9): R137.

Chapter 2 - New Tools for Computational Biologists

2.1 The Main Challenge for a Systems Biologist: Huge and Heterogeneous Data

We are fortunate to live in an era when many functional genomics questions can be answered with high throughput sequencing experiments. This is typically done in a two-step process. First, a researcher applies an experimental technique to isolate particular segments of the genome such as those bound to a protein of interest (ChIP-Seq), or open/accessible (DNase-Seq, ATAC-Seq), or bound to a distal part of the genome (Hi-C methods). Next, these genomic fragments are sequenced and mapped to the genome, and the researcher can infer from the location and amount of reads an answer to the functional question.

At first glance, it seems like an easy task to analyze high-throughput data, since in each case we align reads to a reference genome (or transcriptome for RNA-Seq). After alignment, though, the course of analysis diverges rapidly depending on the particular protocol and research question. ATAC-Seq reads must be shifted an appropriate number of basepairs to account for the location of the Tn5 transposase, aligned ChIP-Seq reads must be compared to the distribution of reads in a control sample to separate true binding from noise, and Hi-ChIP output needs to be binned into sections on the genome before interactions can be identified. Additional computational methods such as identifying transcription factor footprints for hundreds of motifs, or differential expression analysis of different pathways introduce further complexity.

When a research project involves one type of *Seq data and the data are received all at once, this is manageable without any additional tools. However, in the course of a typical systems biology project, a computational biologist will receive data from several different kinds of *Seq experiments, in a set of batches across time, and often the computational biologist is working on other projects simultaneously. In order to effectively meet these challenges, I developed SeqTools, an object-oriented system for managing scientific projects with high-throughput data, SeqReports, a set of analysis tools for differential accessibility / differential expression, and Overlaps3 for identifying enrichment of particular genomic features in a set of loci.

2.2 SeqTools: An Object-Oriented System for *Seq Datasets.

Peter Van Roy defines a programming paradigm as “an approach to programming a computer based on a mathematical theory or a coherent set of principles” (Van Roy, 2009). The object-oriented programming (OOP) paradigm emphasizes organization: the programmer designs objects (data structures) to best encapsulate the complex data he/she is working with and then designs functions to act on these objects. That description is too abstract to be useful, but will become clearer when we look at some concrete examples.

As noted in the last section, all of our experiments involve reads mapped to the genome. Whether it is ATAC-Seq, ChIP-Seq, or RNA-Seq, all of these involve getting a set of fastq files from the sequencer, aligning it to the genome, and then carrying out set of methods on it. We can take advantage of that generality, and consider each of these sequencing datasets members of a **class,** a class I call "SeqData". We can think of all the attributes that a particular instance/object of SeqData might have, such as "single or paired end", "raw fastq files", "reference genome", "aligned output file",etc. and make sure that the SeqData object has a slot for each of these values. With all of this information organized in one object, we can easily access attributes of the dataset by typing in something like:

```
sdSample1.experiment_type
sdSample1.rawFastq1_file_one
```

We can also think of things we would like to do with that particular instance of SeqData, like align it to the genome. Since all SeqData objects have the same structure, we can write an alignment function that looks in the object and retrieves all of the information it needs. Thus, instead of many lines of code specifying what to do, we can just write one:

```
Align_And_CheckQuality(sdSample1)
```

The advantage of all this is that, in exchange for a great deal of initial effort in carefully designing the objects and functions, we can very easily apply functions like "Align_And_CheckQuality" to these objects without having to waste time coding all the specific details like the location of the reference genome, etc. This also standardizes the output from these functions, and makes it much easier to operate on them systematically.

I have also integrated this system with Python doit (http://pydoit.org/) , an automation tool designed by Eduardo Schettino. Doit allows users to specify a project as a series of tasks where the output of one is the input to another. (Say, "Alignment_To_Genome" → "Call Peaks") If a user has a complicated ten step process, and he/she wants to make a change at step five, the package knows to only run the steps downstream of that (steps six to ten), and not waste time on anything upstream. Furthermore, if say, new ATAC-Seq samples are added and the user has already aligned the last batch, the doit tool will see that the output already exists for the old samples and will not rerun the pipeline on them.

2.3 SeqReports and Chromatin Viewer

One of the challenges in a typical systems biology project is simply organizing all of the output from routine analyses such as checking the quality of a given batch of samples or finding differentially expressed genes. These results are often wind up dispersed in a

collection of slide decks, pdfs, and emails, making it difficult to track down results in the future. To combat this tendency, I have created a web-based view of the quality information on our ChIP-Seq and ATAC-Seq data using "Shiny", and also have written our differential expression / differential peak reports as knitr html reports for collaborators.

Shiny ("Web Application Framework for R [R Package Shiny Version 1.0.5]" n.d.) is a package for the R programming language that makes it easy to build interactive websites. Using the SeqTools package, I systematically assemble various quality control metrics on our ATAC-Seq and ChIP-Seq results into tables of results that can be loaded in Shiny and plotted with R. Since SeqTools automatically updates these tables every time it aligns new data, I merely have to restart Shiny to get an updated version of the viewer that plots all of our samples from beginning to end. The bars for individual samples are also colored by batch. This makes it easy to quickly compare a new set of results to previous ones, switch through many different metrics of quality, and thus identify problematic samples.

In addition to Shiny, I also use the knitr package (Xie 2015) to produce reports that combine text and plots from a differential ATAC-seq / RNA-seq analysis into a single html report. Included with the report are descriptions of the motivation and mechanics for different results we plot (such as how the Euclidean distance for a heatmap is calculated, and how PCA works), to help explain the methods I used. For analysis of differential results in ATAC-Seq, I also incorporate analysis from "Genomic Regions Enrichment of Annotations Tool" (GREAT) (McLean et al. 2010) , which maps regulatory regions to their putative targets. The knitr program will upload these peaks to GREAT's API and also prepare peaks for viewing in the Integrated Genomic Viewer (IGV).

2.4 Overlaps3

After alignment and peak-calling of ATAC-Seq or ChIP-Seq data, we have a set of loci on the genome and want to know if there is any evidence that the loci are associated with regulatory functions. There are several pieces of evidence we can consider for a particular set of loci, such as their distance to transcription start sites (TSS's), overlap with known collections of regulatory regions, and enrichment for evolutionarily conserved regions and other features. To systematically check for enrichment with these features, I wrote the Overlaps3 package for R.

Overlaps3 can run in two ways: 1) to check a set of loci on the genome for enrichment of features relative to the full genome, or 2) to compare a "foreground set" of loci for enrichment of features relative to a "background set" of loci. To clarify the differences between the two methods, imagine you have performed ATAC-Seq on T cells in two different conditions. You may want to take the union of all peaks, and run it through method #1 to confirm that you are finding features associated with T cells in general (say, motifs of key TFs, T cell pathways in GREAT, etc.) You then may want to drill down

to peaks that are differentially closed/open between the two conditions and try to identify what is particularly enriched in them compared to your other peaks. If you run this narrow set of loci through method #1, you will get many hits for T cell pathways, most of which overlap with your previous results. However, if you use method #2 with the differential peaks as the foreground and all T cell peaks as the background, you will limit your results to those that distinguish the differential peaks, and these will provide a better understanding of their particular function.

For method #2, we measure enrichment for an annotation as the percent of loci in the foreground set that overlap a feature of interest, divided by the percent of loci in the background set that overlap the same feature. In the example above, if 35% of the foreground loci had a hit for a given motif, and 10% of the background regions also had a hit, the enrichment score would be 3.5 . We use a hypergeometric test to assess the significance of these results.

For method #1, we want to know how enriched the loci are for particular features relative to a set of loci of the same size randomly chosen from the genome. We use the regioneR package (Gel et al. 2016) to randomly draw regions of the same size as our loci from the genome, count how many of these loci overlap features of interest, and repeat this hundreds of times to get an estimate from the genome. For method #1, if 30 % of our loci overlapped “Feature A”, and the average overlap for 500 random draws of regions the same size was 10% for “Feature A”, the enrichment score would be 3.0 . We use the mean and standard deviation of these enrichment scores of these draws to parameterize a normal distribution, which is used to compute a p value for our enrichment score.

The Overlaps3 package provides four kinds of output: 1) Enrichment score for motifs and genomic annotations, 2) Enrichment score for disease SNPs obtained from the package “gwascat” Carey V (2018) [mm10 loci are mapped to hg38 using the liftOver package from UCSC and Bioconductor], 3) GREAT results (McLean et al. 2010), and 4) Enrichment for proximity to lists of user-supplied genes. (The peak-to-gene mapping is performed by GREAT.)

2.5 References

Carey V. 2018. gwascat: representing and modeling data in the EMBL-EBI GWAS catalog. R package version 2.12.0.

Gel, Bernat, Anna Díez-Villanueva, Eduard Serra, Marcus Buschbeck, Miguel A. Peinado, and Roberto Malinverni. 2016. “regioneR: An R/Bioconductor Package for the Association Analysis of Genomic Regions Based on Permutation Tests.” *Bioinformatics* 32 (2): 289–91.

McLean, Cory Y., Dave Bristor, Michael Hiller, Shoa L. Clarke, Bruce T. Schaar, Craig B. Lowe, Aaron M. Wenger, and Gill Bejerano. 2010. “GREAT Improves Functional Interpretation of Cis-Regulatory Regions.” *Nature Biotechnology* 28 (5): 495–501.

Van Roy, P. n.d. “Programming Paradigms for Dummies: What Every Programmer Should Know.” 2009. In G. Assayag and A. Gerzso (eds.) New Computational Paradigms for Computer Music, IRCAM/Delatour, France.
Xie, Yihui. 2015. *Dynamic Documents with R and Knitr, Second Edition (Chapman & Hall/CRC The R Series)*. 2 edition. Chapman and Hall/CRC.
“Web Application Framework for R [R Package Shiny Version 1.0.5].” n.d. Accessed March 5, 2018. https://cran.r-project.org/web/packages/shiny/index.html.

Chapter 3 - The Epigenetic Landscape of T Cell Exhaustion

3.1 Introduction

My first research project on the regulation of CD8+ T Cell fate began by analyzing ATAC-Seq data from Naive, Effector, Memory, and Exhausted T Cells. I used the SeqTools infrastructure, Chromatin Viewer, and an early version of the Overlaps3 database ("OverlapsDB") to process and analyze the data. My co-authors and I were able to find chromatin accessible regions (ChAR's) that were specific to the exhausted conditions, and also validated one of them as an enhancer for PD-1, an important cell surface protein that is upregulated during T Cell exhaustion and is a target of immunotherapy.

This paper was the work of many people, and I am reprinting it here as it was published in *Science* in 2016 (Sen et al. 2016). The authors on the paper are:

Debattama R. Sen[1,2,*], James Kaminski[3,*], R. Anthony Barnitz[1], Makoto Kurachi[4,5], Ulrike Gerdemann[1], Kathleen B. Yates[1], Hsiao-Wei Tsao[1], Jernej Godec[1,2], Martin W. LaFleur[1,2], Flavian D. Brown[1,2], Pierre Tonnerre[6], Raymond T. Chung[6], Damien C. Tully[7], Todd M. Allen[7], Nicole Frahm[8], Georg M. Lauer[6], E. John Wherry[4,5], Nir Yosef[3,7,9,†,‡], W. Nicholas Haining[1,10,11,†,‡]

[1] Department of Pediatric Oncology, Dana-Farber Cancer Institute, Boston, MA 02115, USA.
[2] Division of Medical Sciences, Harvard Medical School, Boston, MA 02115, USA.
[3] Center for Computational Biology, University of California, Berkeley, Berkeley, CA 94720, USA.
[4] Institute of Immunology, University of Pennsylvania, Philadelphia, PA 19104, USA.
[5] Department of Microbiology, University of Pennsylvania, Philadelphia, PA 19104, USA.
[6] Gastrointestinal Unit and Liver Center, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02115, USA.
[7] Ragon Institute of Massachusetts General Hospital, Massachusetts Institute of Technology, and Harvard University, Boston, MA 02139, USA.
[8] Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA.
[9] Department of Electrical Engineering and Computer Science, University of California, Berkeley, Berkeley, CA 94720, USA.
[10] Division of Pediatric Hematology and Oncology, Children's Hospital, Boston, MA 02115, USA.
[11] Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, MA 02142, USA.
[‡]Corresponding author. Email: niryosef@berkeley.edu (N.Y.); nicholas_haining@dfci.harvard.edu (W.N.H.)
[*] These authors contributed equally to this work.
[†] These authors contributed equally to this work.

3.2 Abstract

Exhausted T cells in cancer and chronic viral infection express distinctive patterns of genes, including sustained expression of programmed cell death protein 1 (PD-1). However, the regulation of gene expression in exhausted T cells is poorly understood. Here, we define the accessible chromatin landscape in exhausted CD8+ T cells and show that it is distinct from functional memory CD8+ T cells. Exhausted CD8+ T cells in humans and a mouse model of chronic viral infection acquire a state-specific epigenetic landscape organized into functional modules of enhancers. Genome editing shows that PD-1 expression is regulated in part by an exhaustion-specific enhancer that contains essential RAR, T-bet, and Sox3 motifs. Functional enhancer maps may offer targets for genome editing that alter gene expression preferentially in exhausted CD8+ T cells.

3.3 The Epigenetic Landscape of T Cell Exhaustion

T cell exhaustion—an acquired state of T cell dysfunction—is a hallmark of cancer and chronic viral infection (1, 2), and clinical trials of checkpoint blockade, which aim to reverse T cell exhaustion in cancer, have proven strikingly effective (3, 4). Chimeric antigen receptor (CAR)–T cell therapy has also proven highly effective for hematologic malignancies (5), but the development of exhaustion in T cells engineered to treat solid tumors remains a substantial barrier to its broader use (6). The identification of mechanisms that regulate exhausted T cells is therefore a major goal in cancer immunotherapy.

To identify regulatory regions in the genome of exhausted CD8+ T cells, we used an assay for transposase-accessible chromatin with high-throughput sequencing (ATAC-seq) (7) to demarcate areas of accessible chromatin in mouse antigen-specific CD8+ T cells differentiating in response to lymphocytic choriomeningitis virus (LCMV) infection (fig. S1A and table S1). Acute LCMV infection elicits highly functional effector CD8+ T cells, whereas chronic LCMV infection gives rise to exhausted CD8+ T cells (1–3, 8, 9). Analysis of high-quality ATAC-seq profiles (fig. S1, B to H) from naïve CD8+ T cells and those at day 8 and day 27 postinfection (p.i.) (d8 and d27, respectively) revealed that naïve CD8+ T cells underwent large-scale remodeling (Fig. 1A and fig. S2A) during differentiation [as detected by DESeq2, with a false discovery rate (FDR) < 0.05]. The majority (71%) (fig. S2A) of chromatin-accessible regions (ChARs) either emerged (e.g., those at the Ifng locus) or disappeared (e.g., Ccr7) (Fig. 1A) as naïve CD8+ T cells underwent differentiation. The gain and loss of ChARs were not balanced; a much larger fraction of regions emerged at d8 p.i. and persisted or emerged only at d27 than were either transiently detected at d8 p.i. or lost from naïve cells (Fig. 1B). Thus, differentiation from a naïve CD8+ T cell state is associated with a net increase, rather than decrease, in chromatin accessibility (fig. S2B).

Genes adjacent to ChARs in each state-specific module included many with known functions in the corresponding T cell state. For example, module d, active in mouse T cells experiencing chronic LCMV infection on d8 and d27 p.i., contained ChARs adjacent to the inhibitory receptors Pdcd1 and Havcr2 (which encodes Tim3) and the transcription factor Batf, all genes that are up-regulated in exhausted CD8+ T cells (Fig. 2B) (1, 8). Moreover, the functional classes of genes in each module were distinct on the basis of pathway enrichment (Fig. 2C and table S2). Thus, ChARs that distinguish naïve, effector, memory, and exhausted CD8+ T cells are organized into state-specific modules that positively regulate functionally distinct programs of genes.

We next sought to test whether regulatory regions specific to exhausted cells could regulate genes differentially expressed in exhausted CD8+ T cells. Persistent expression of PD-1 is a cardinal feature of exhausted CD8+ T cells, but PD-1 is also transiently expressed by effector CD8+ T cells during acute LCMV infection (3, 8). We identified nine ChARs within 45 kb of the Pdcd1 gene locus (Fig. 3A) and found several that correspond to previously described regions with enhancer activity (–1.5 kb and –3.7 kb) (Fig. 3A) (10); these were present in both acute and chronic infection. We also identified an additional region (–23.8 kb) that only showed appreciable chromatin accessibility in exhausted CD8+ T cells at d8 and d27 p.i. from chronic infection (Fig. 3A).

We hypothesized that this ChAR might function as an enhancer of PD-1 that is required for persistent, high levels of expression in exhausted CD8+ T cells. Analysis of chromatin accessibility at this region in previously published deoxyribonuclease I–hypersensitive site–mapping (11) or ATAC-seq data (12) showed that it was not active in other types of hematopoietic cells, except the murine T cell line EL4 and regulatory CD4+ T cells, both of which can constitutively express high levels of PD-1 (10, 13) (Fig. 3A and fig. S4A). We cloned a 781–base pair (bp) fragment corresponding to this region into a reporter construct and found that it induced a 10- to 12-fold increase in reporter gene expression, confirming that it could function as an enhancer (fig. S4B).

We then tested whether the –23.8 kb enhancer was necessary for high-level PD-1 expression. We used the CRISPR-Cas9 nuclease to delete a 1.2-kb fragment at that position in EL4 cells, which have both sustained high-level PD-1 expression and open chromatin at that enhancer site (14, 15) (fig. S4, C to G). In Cas9-expressing EL4 cells transduced with a pair of single-guide RNAs (sgRNAs) flanking the enhancer, cells with the lowest PD-1 expression had the highest amount of the enhancer deletion (Fig. 3B). We confirmed this finding in single-cell clones and found that the expression of PD-1 in clones with a biallelic deletion of the target ChAR was significantly lower ($P > 0.0002$, Mann-Whitney U test) than expression in nondeleted clones (fig. S4, H to J). Deletion of this region resulted in decreased but not abrogated PD-1, suggesting that additional regulatory regions in EL4 cells are also involved in regulating PD-1 expression (Fig. 3C). Among all genes within 1.5 Mb of the Pdcd1 locus, only PD-1 mRNA expression was significantly decreased by deletion of the –23.8 kb ChAR (fig. S3K). This suggests that the

–23.8 kb ChAR present in exhausted, but not functional, CD8+ T cells serves as an enhancer that is required to maintain high levels of PD-1 expression.

We next sought to identify the functional contribution of specific sequences within enhancer regions to the regulation of PD-1 expression. We used Cas9-mediated in situ saturation mutagenesis and designed all possible sgRNAs within the –23.8 kb enhancer and eight other regulatory sequences near the Pdcd1 locus (15, 16) (Fig. 3A). We transduced Cas9-expressing EL4 cells with a pool of 1754 enhancer-targeting sgRNAs, 117 sgRNAs targeting the Pdcd1 exons as positive controls, and 200 nontargeting sgRNAs as negative controls (fig. S5, A and B). We sorted transduced EL4s into populations on the basis of high or low PD-1 expression and quantified the abundance of individual sgRNAs (fig. S5C).

In comparison with nontargeting sgRNAs, which were equivalently distributed between PD-1–high and PD-1–low fractions, sgRNAs targeting Pdcd1 exons were highly enriched in the PD-1–low fraction as expected (Fig. 3D and fig. S5, D and E). sgRNAs targeting eight of the nine regulatory regions were also significantly enriched in the PD-1–low fraction to varying degrees ($P < 0.00001$ to $P < 0.01$, see supplementary methods), suggesting that critical sequences affecting PD-1 expression are densely represented within each of the eight regulatory regions. However, sgRNAs in the –35.6 kb ChAR had no significant effect on PD-1 expression, consistent with prior observations that this region falls outside the CCCTC-binding factor (CTCF)–mediated boundaries of the Pdcd1 locus (10).

We focused on sgRNAs inducing cleavage in the –23.8 kb enhancer (fig. S5F) and found a strong correlation between the predicted activity in a pooled setting (PD-1 high:low ratio >1 SD below mean) and their effect on PD-1 mean fluorescence intensity in individual cell lines ($P = 0.0041$) (fig. S5, G and H). Inspection of the predicted cleavage-site locations revealed three critical regions of the enhancer in which cleavage markedly affected PD-1 expression (Fig. 3E, gray shading).

We next asked whether these critical regions in the –23.8 kb enhancer were associated with distinct patterns of transcription factor (TF) binding in exhausted CD8+ T cells in vivo. We identified TF footprints (17) using ATAC-seq cut sites from CD8+ T cells experiencing chronic infection, which allowed us to infer TF binding within the –23.8 kb enhancer (Fig. 3E; fig. S6, A to D; fig. S7A; and tables S3 to S6). We found that cleavage sites of sgRNAs that reduced PD-1 expression in EL4 cells were significantly enriched in TF footprints found in exhausted CD8+ cells in vivo ($P = 8.63 \times 104$, hypergeometric test). The three TF footprints with greatest sensitivity to disruption corresponded to motifs for Sox3, T-bet (encoded by Tbx21), and retinoic acid receptor (RAR) in exhausted CD8+ T cells in vivo (Fig. 3F and fig. S7B). Indeed, comparison of genome-wide TF footprinting between chronic and acute infection at d27 to identify TF motifs that showed significantly differential inferred binding (Fig. 3G, fig. S7C, and tables S3 and S5)

confirmed that Rara binding was significantly enriched in exhausted CD8+ T cells (FDR = 3.14 × 10−13) compared with their functional counterparts.

To test whether T cell exhaustion is also associated with a distinct epigenetic state in human exhausted CD8+ T cells, we analyzed global patterns of chromatin accessibility in tetramer+ CD8+ T cells from four subjects with chronic progressive HIV-1 who were not on therapy (Fig. 4, A and B; fig. S8, A and B; and table S7). We successfully mapped 80 to 85% of ChARs identified in the mouse model to their human orthologous regions (Fig. 4A, colored blocks, and fig. S8C) (18, 19) and found them to be enriched for disease-associated single-nucleotide polymorphisms (SNPs) (probabilistic identification of causal SNPs, $P < 2.77 \times 10{-}8$; hypergeometric test) (fig. S8D) (20) and, in particular, immune-related National Human Genome Research Institute genome-wide association study SNPs ($P < 3.70 \times 10{-}3$) (fig. S8, E to G). This enrichment strongly suggested that mapped regions corresponded to functional regulatory regions within the immune system. Regions at the Pdcd1 locus were not among those mapped from the mouse model, as previously observed (10), which limited our ability to detect an ortholog to the –23.8 kb enhancer observed in the mouse model.

Human naïve CD8+ T cells from the majority of donors showed greater chromatin accessibility in naïve-specific regions defined in the mouse than in memory- or exhaustion-specific regions. In the healthy donor, CMV-specific tetramer+ CD8+ T cells, and effector memory cells were enriched for memory-specific regions (Mann-Whitney U test, $P = 0.01$ to $P < 0.0001$) (Fig. 4C and fig. S8H). In contrast, HIV-specific tetramer+ cells from three out of the four subjects showed significantly greater chromatin accessibility in exhaustion-specific regions (Mann-Whitney U test, $P = 0.05$ to $P < 0.001$) than in memory-specific regions.

Finally, we confirmed these findings in a subject with chronic hepatitis C virus (HCV) infection in whom CD8+ T cell responses to two epitopes of HCV could be detected (Fig. 4D). Sequencing of the HCV genome in this subject revealed that, unlike the C63B epitope, the 174D epitope had undergone extensive viral escape, and no wild-type viral sequence could be detected (Fig. 4E). We found that the C63B tetramer+ cells had a phenotype consistent with exhaustion and showed significantly greater chromatin accessibility at exhaustion-specific regions (Mann-Whitney U test, $P = 0.01$) than memory regions (Fig. 4F). In contrast, 174D tetramer+ cells, which were specific for the escape mutant epitope, lacked exhaustion-specific surface markers and showed greater chromatin accessibility in memory-specific regions, as did influenza-specific CD8+ T cells (Mann-Whitney U test, $P = 0.04$) (Fig. 4F). Thus, the state-specific pattern of chromatin accessibility found in mouse exhausted CD8+ T cells is conserved in human exhausted CD8+ T cells.

We find that CD8+ T cell exhaustion occurs with a broad remodeling of the enhancer landscape and TF binding. This suggests that exhausted CD8+ T cells occupy a differentiation state distinct from functional memory CD8+ T cells. Identifying the

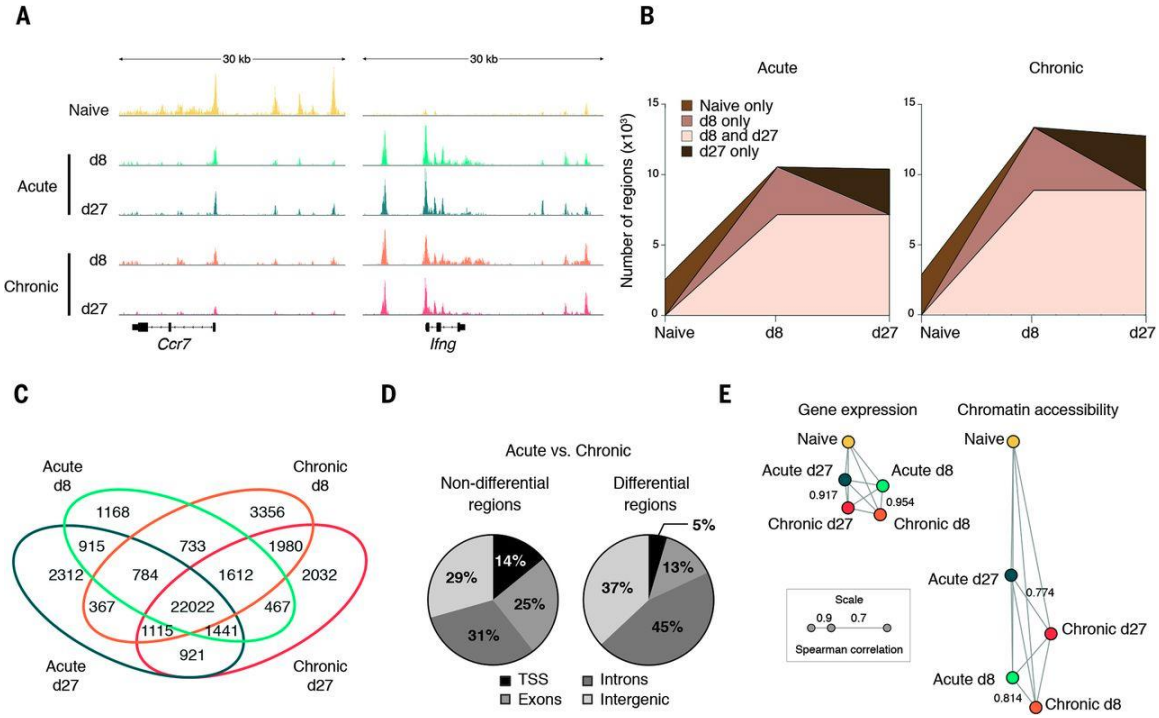
plasticity of this state and whether or how it could be reverted becomes a critical question for immunotherapy applications. Our data also suggest that mapping state-specific enhancers in exhausted T cells could enable more precise genome editing for adoptive T cell therapy. Genome editing of CAR-T cells to make them resistant to exhaustion is an appealing concept and has led to recent studies investigating the deletion of the PD-1 gene locus (21, 22). Editing exhaustion-specific enhancers (15) may provide a more “tunable” and state-specific approach to modulate T cell function than deleting coding regions of genes. Functional maps of enhancers specific to exhausted CD8+ T cells may therefore provide a crucial step toward the rational engineering of T cells for therapeutic use.

3.4 Acknowledgements

The authors thank members of the Yosef and Haining laboratories for their input, the research subjects for their participation, and J. Doench and the entire Genetic Perturbation Platform at the Broad Institute for their advice on Cas9-mediated screening technology. The authors are grateful for input from the Cancer Center for Genome Discovery. The data reported in this manuscript are tabulated in the main paper and in the supplementary materials. Genome-wide data generated in this study can be accessed via GEO accession no. GSE87646. This research was supported by AI115712, AI091493, and AI082630 to W.N.H. from the NIH; by the BRAIN Initiative grants MH105979 and HG007910 from the NIH to N.Y.; and by 1R21AI078809-01 and UM1 AI068618 from the NIH to N.F. The authors declare no potential conflicts of interest. D.R.S., J.K., N.Y., E.J.W., and W.N.H. are inventors on a patent application (U.S. Patent Application no. 62/310,903) held and submitted by Dana-Farber Cancer Institute.

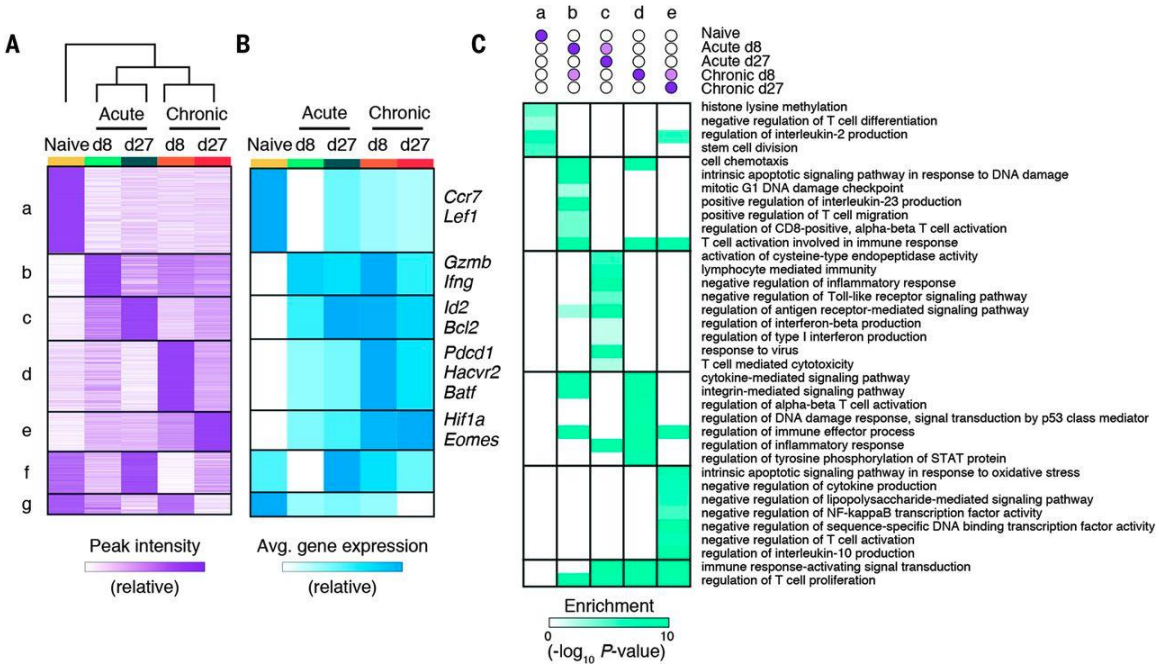
3.5 Figures

Figure 1 - CD8+ T cell exhaustion is associated with extensive changes in accessible chromatin.



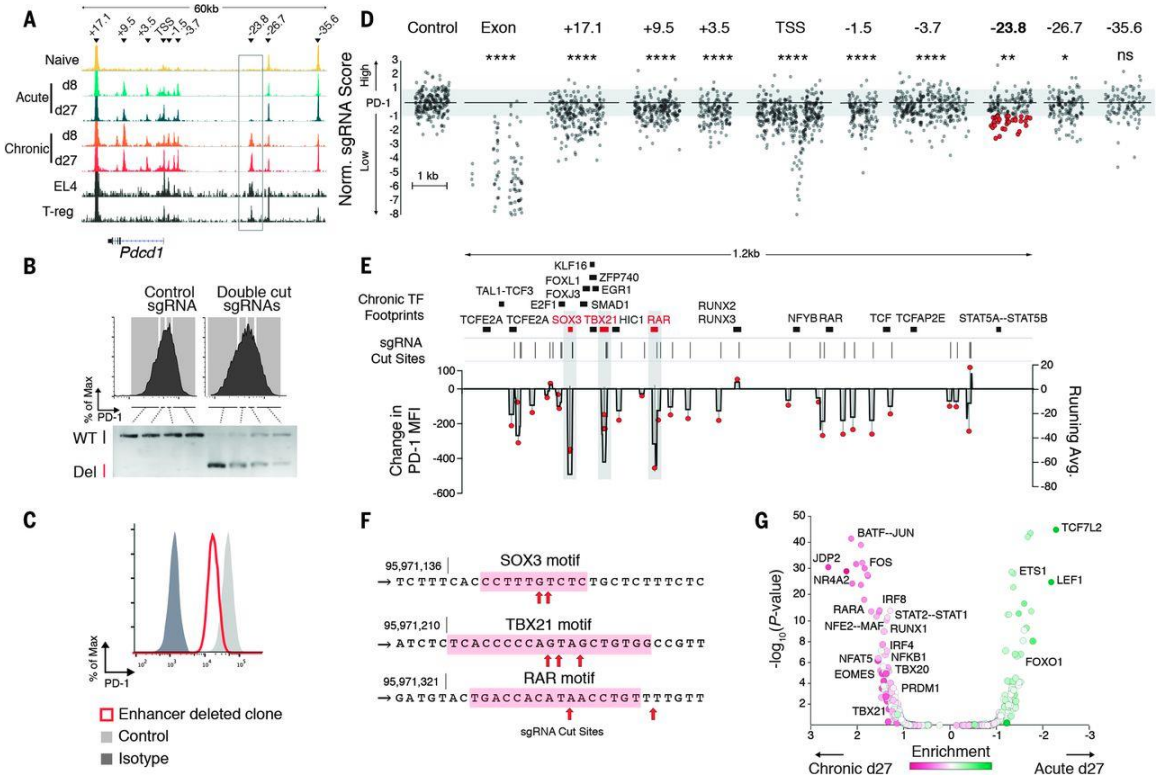
(A) Representative ATAC-seq tracks at the *Ccr7* and *Ifng* gene loci. (B) Developmental trajectory of new regions at each time point. (C) Overlap in ChARs between cell states. (D) Distribution of nondifferential (left) and differential (right) regions between acute and chronic CD8⁺ T cell states. TSS, transcription start site. (E) Correlation network of similarity between states measured by gene expression (left) and chromatin accessibility (right). Edge length corresponds to similarity (Spearman correlation).

Figure 2 - State-specific enhancers in CD8+ T cells form modules that map to functionally distinct classes of genes.



(A) Heat map of peak intensity for all differentially accessible regions (rows) clustered by similarity across cell states (columns). Shown are normalized numbers of cut sites (supplementary methods), scaled linearly from row minimum (white) to maximum (purple). (B) Heat map showing row-normalized average mRNA expression of neighboring genes within each module in (A) in each cell state. Informative genes from each module are shown on right. (C) Heat map showing enrichment of Gene Ontology (GO) terms (rows) in each module (columns). *P*-values (hypergeometric test) presented as $-\log_{10}$.

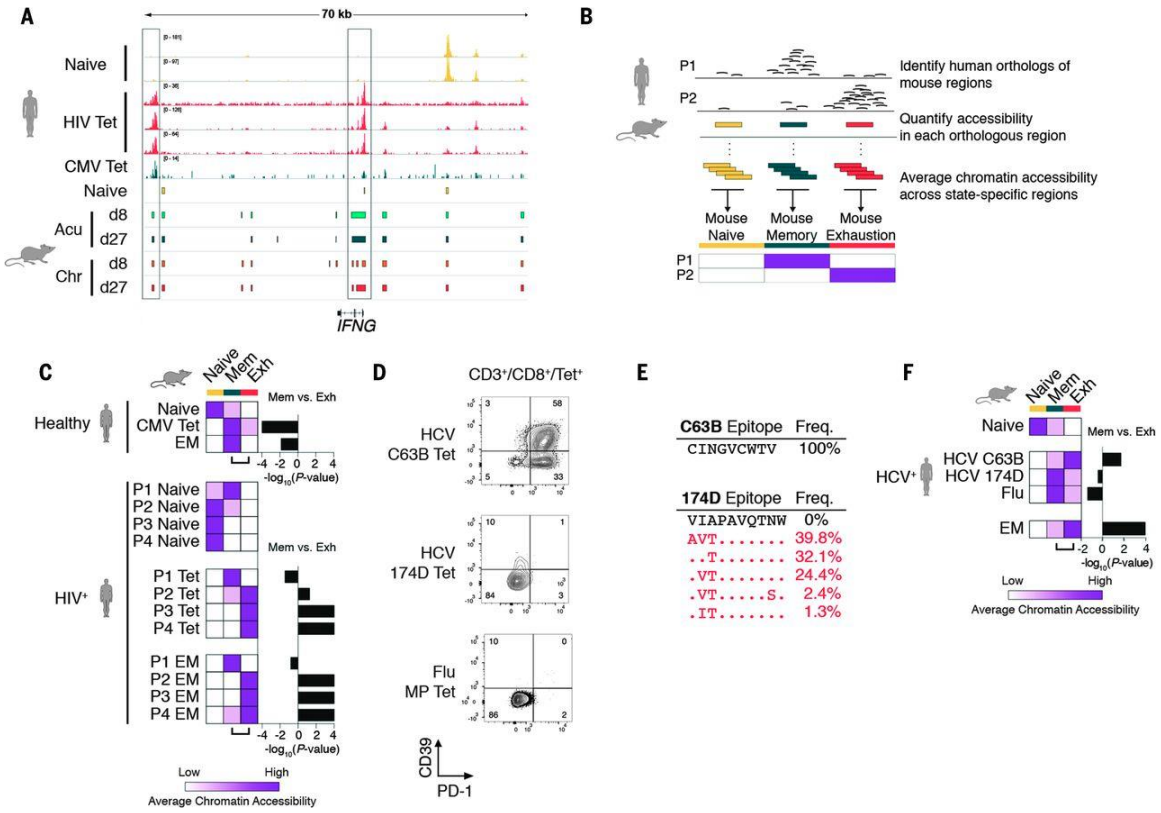
Figure 3 - High-resolution functional mapping of an exhaustion-specific enhancer identifies minimal sequences that regulate PD-1.



(A) ATAC-seq tracks from CD8$^+$ T cells, EL4 cell line, and regulatory CD4$^+$ T cells (*12*). Arrowheads indicate individual ChARs. (B) Cell sorting gates (top) and corresponding genomic polymerase chain reaction amplification for the PD-1 enhancer region (bottom) showing proportion of wild-type (WT) or deleted (Del) alleles in EL4 cells transfected with control (left) or double-cut sgRNAs (right). Representative data shown from two replicates. (C) PD-1 expression of EL4 WT (light gray) or representative enhancer-deleted (red) single-cell clone out of 46 clones. (D) Normalized enrichment of sgRNAs (gray symbols) within PD-1–high and PD-1–low populations at locations shown (supplementary methods). Control nontargeting sgRNAs are pseudo mapped with 5-bp spacing. Red symbols correspond to the 21 sgRNAs with the largest effect within the –23.8 kb enhancer, for which isogenic cell lines were later produced. (E) Overlap of TF footprints and sgRNA activity within the –23.8 kb enhancer. TF footprints with binding probability >0.9 in chronic d27 are shown on top. Lines represent cut sites of top-scoring sgRNAs. Change in PD-1 mean fluorescence intensity (MFI) relative to control guide transfected populations for each sgRNA (red symbol, left axis); 10-bp running average of PD-1 MFI changes caused by sgRNA activity shown in black (right axis). (F) sgRNA cut sites within the SOX3, TBX21, and RAR motifs. (G) Log fold enrichment of predicted TF

footprints in acute d27 versus chronic d27 $CD8^+$ T cells (*x* axis) (see supplementary methods) are plotted against the corresponding *P*-value (hypergeometric).

Figure 4 - Exhaustion-specific epigenetic profiles in the mouse are conserved in antigen-specific exhausted human T cells in HIV-1 infection.



(A) Representative ATAC-seq tracks from naïve, HIV-1 tetramer$^+$, and CMV tetramer$^+$ samples at the *IFNG* gene locus (top). Orthologous regions from five mouse cell states at the *IFNG* locus, based on mapping of mouse ChARs to the human genome (colored blocks, bottom). (B) Schematic diagram of mouse and human comparative analysis. (C) Heat map of average chromatin accessibility at regions orthologous to mouse naïve, memory, and exhaustion enhancers in human samples indicated. Color scale as in Fig. 2A. (D) PD-1 and CD39 expression measured by flow cytometry in HCV C63B tetramer$^+$, HCV 174D tetramer$^+$, and influenza (flu) matrix peptide (MP) tetramer$^+$ populations from a single HCV-infected donor. (E) Viral sequences encoding C63B and 174D epitopes. (F) Heat map of average chromatin accessibility at regions orthologous to mouse naïve, memory, and exhaustion enhancers in human samples indicated from a single HCV-infected donor.

3.6 References

1 . E. J. Wherry, S.-J. J. Ha, S. M. Kaech, W. N. Haining, S. Sarkar, V. Kalia, S. Subramaniam, J. N. Blattman, D. L. Barber, R. Ahmed, Molecular signature of CD8+ T cell exhaustion during chronic viral infection. Immunity 27, 670–684 (2007). doi:10.1016/j.immuni.2007.09.006pmid:17950003

2. A. J. Zajac, J. N. Blattman, K. Murali-Krishna, D. J. Sourdive, M. Suresh, J. D. Altman, R. Ahmed, Viral immune evasion due to persistence of activated T cells without effector function. J. Exp. Med. 188, 2205–2213 (1998). doi:10.1084/jem.188.12.2205pmid:9858507

3. D. L. Barber, E. J. Wherry, D. Masopust, B. Zhu, J. P. Allison, A. H. Sharpe, G. J. Freeman, R. Ahmed, Restoring function in exhausted CD8 T cells during chronic viral infection. Nature 439, 682–687 (2006). doi:10.1038/nature04444pmid:16382236

4. S. L. Topalian, F. S. Hodi, J. R. Brahmer, S. N. Gettinger, D. C. Smith, D. F. McDermott, J. D. Powderly, R. D. Carvajal, J. A. Sosman, M. B. Atkins, P. D. Leming, D. R. Spigel, S. J. Antonia, L. Horn, C. G. Drake, D. M. Pardoll, L. Chen, W. H. Sharfman, R. A. Anders, J. M. Taube, T. L. McMiller, H. Xu, A. J. Korman, M. Jure-Kunkel, S. Agrawal, D. McDonald, G. D. Kollia, A. Gupta, J. M. Wigginton, M. Sznol, Safety, activity, and immune correlates of anti-PD-1 antibody in cancer. N. Engl. J. Med. 366, 2443–2454 (2012). doi:10.1056/NEJMoa1200690pmid:22658127

5. D. L. Porter, B. L. Levine, M. Kalos, A. Bagg, C. H. June, Chimeric antigen receptor-modified T cells in chronic lymphoid leukemia. N. Engl. J. Med. 365, 725–733 (2011). doi:10.1056/NEJMoa1103849pmid:21830940

6. A. H. Long, W. M. Haso, J. F. Shern, K. M. Wanhainen, M. Murgai, M. Ingaramo, J. P. Smith, A. J. Walker, M. E. Kohler, V. R. Venkateshwara, R. N. Kaplan, G. H. Patterson, T. J. Fry, R. J. Orentas, C. L. Mackall, 4-1BB costimulation ameliorates T cell exhaustion induced by tonic signaling of chimeric antigen receptors. Nat. Med. 21, 581–590 (2015). doi:10.1038/nm.3838pmid:25939063

7. J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, W. J. Greenleaf, Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods 10, 1213–1218 (2013). doi:10.1038/nmeth.2688pmid:24097267

8. T. A. Doering, A. Crawford, J. M. Angelosanto, M. A. Paley, C. G. Ziegler, E. J. Wherry, Network analysis reveals centrally connected genes and pathways involved in CD8+ T cell exhaustion versus memory. Immunity 37, 1130–1144 (2012).

doi:10.1016/j.immuni.2012.08.021pmid:23159438

9. M. A. Paley, D. C. Kroy, P. M. Odorizzi, J. B. Johnnidis, D. V. Dolfi, B. E. Barnett, E. K. Bikoff, E. J. Robertson, G. M. Lauer, S. L. Reiner, E. J. Wherry, Progenitor and terminal subsets of CD8+ T cells cooperate to contain chronic viral infection. Science 338, 1220–1225 (2012). doi:10.1126/science.1229620pmid:23197535

10. J. W. Austin, P. Lu, P. Majumder, R. Ahmed, J. M. Boss, STAT3, STAT4, NFATc1, and CTCF regulate PD-1 through multiple novel regulatory regions in murine T cells. J. Immunol. 192, 4876–4886 (2014). doi:10.4049/jimmunol.1302750pmid:24711622

11. J. Vierstra, E. Rynes, R. Sandstrom, M. Zhang, T. Canfield, R. S. Hansen, S. Stehling-Sun, P. J. Sabo, R. Byron, R. Humbert, R. E. Thurman, A. K. Johnson, S. Vong, K. Lee, D. Bates, F. Neri, M. Diegel, E. Giste, E. Haugen, D. Dunn, M. S. Wilken, S. Josefowicz, R. Samstein, K.-H. H. Chang, E. E. Eichler, M. De Bruijn, T. A. Reh, A. Skoultchi, A. Rudensky, S. H. Orkin, T. Papayannopoulou, P. M. Treuting, L. Selleri, R. Kaul, M. Groudine, M. A. Bender, J. A. Stamatoyannopoulos, Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. Science 346, 1007–1012 (2014). doi:10.1126/science.1246426pmid:25411453

12. D. Lara-Astiaso, A. Weiner, E. Lorenzo-Vivas, I. Zaretsky, D. A. Jaitin, E. David, H. Keren-Shaul, A. Mildner, D. Winter, S. Jung, N. Friedman, I. Amit, Chromatin state dynamics during blood formation. Science 345, 943–949 (2014). doi:10.1126/science.1256271pmid:25103404

13. G. Raimondi, W. J. Shufesky, D. Tokita, A. E. Morelli, A. W. Thomson, Regulated compartmentalization of programmed cell death-1 discriminates CD4+CD25+ resting regulatory T cells from activated T cells. J. Immunol. 176, 2808–2816 (2006). doi:10.4049/jimmunol.176.5.2808pmid:16493037

14. O. Shalem, N. E. Sanjana, E. Hartenian, X. Shi, D. A. Scott, T. S. Mikkelsen, D. Heckl, B. L. Ebert, D. E. Root, J. G. Doench, F. Zhang, Genome-scale CRISPR-Cas9 knockout screening in human cells. Science 343, 84–87 (2014). doi:10.1126/science.1247005pmid:24336571

15. M. C. Canver, E. C. Smith, F. Sher, L. Pinello, N. E. Sanjana, O. Shalem, D. D. Chen, P. G. Schupp, D. S. Vinjamur, S. P. Garcia, S. Luc, R. Kurita, Y. Nakamura, Y. Fujiwara, T. Maeda, G.-C. C. Yuan, F. Zhang, S. H. Orkin, D. E. Bauer, BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. Nature 527, 192–197 (2015). doi:10.1038/nature15521pmid:26375006

16. J. Vierstra, A. Reik, K. H. Chang, S. Stehling-Sun, Y. Zhou, S. J. Hinkley, D. E. Paschon, L. Zhang, N. Psatha, Y. R. Bendana, C. M. O'Neil, A. H. Song, A. K. Mich, P. Q. Liu, G. Lee, D. E. Bauer, M. C. Holmes, S. H. Orkin, T. Papayannopoulou, G.

Stamatoyannopoulos, E. J. Rebar, P. D. Gregory, F. D. Urnov, J. A. Stamatoyannopoulos, Functional footprinting of regulatory DNA. Nat. Methods 12, 927–930 (2015). doi:10.1038/nmeth.3554pmid:26322838

17. R. Pique-Regi, J. F. Degner, A. A. Pai, D. J. Gaffney, Y. Gilad, J. K. Pritchard, Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Res. 21, 447–455 (2011). doi:10.1101/gr.112623.110pmid:21106904

18. E. Gjoneska, A. R. Pfenning, H. Mathys, G. Quon, A. Kundaje, L.-H. Tsai, M. Kellis, Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. Nature 518, 365–369 (2015). doi:10.1038/nature14252pmid:25693568

19. D. Villar, C. Berthelot, S. Aldridge, T. F. Rayner, M. Lukk, M. Pignatelli, T. J. Park, R. Deaville, J. T. Erichsen, A. J. Jasinska, J. M. Turner, M. F. Bertelsen, E. P. Murchison, P. Flicek, D. T. Odom, Enhancer evolution across 20 mammalian species. Cell 160, 554–566 (2015). doi:10.1016/j.cell.2015.01.006pmid:25635462

20. K. K. Farh, A. Marson, J. Zhu, M. Kleinewietfeld, W. J. Housley, S. Beik, N. Shoresh, H. Whitton, R. J. Ryan, A. A. Shishkin, M. Hatan, M. J. Carrasco-Alfonso, D. Mayer, C. J. Luckey, N. A. Patsopoulos, P. L. De Jager, V. K. Kuchroo, C. B. Epstein, M. J. Daly, D. A. Hafler, B. E. Bernstein, Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature 518, 337–343 (2015). doi:10.1038/nature13835pmid:25363779

21. A. Hendel, R. O. Bak, J. T. Clark, A. B. Kennedy, D. E. Ryan, S. Roy, I. Steinfeld, B. D. Lunstad, R. J. Kaiser, A. B. Wilkens, R. Bacchetta, A. Tsalenko, D. Dellinger, L. Bruhn, M. H. Porteus, Chemically modified guide RNAs enhance CRISPR-Cas genome editing in human primary cells. Nat. Biotechnol. 33, 985–989 (2015). doi:10.1038/nbt.3290pmid:26121415

22. K. Schumann, S. Lin, E. Boyer, D. R. Simeonov, M. Subramaniam, R. E. Gate, G. E. Haliburton, C. J. Ye, J. A. Bluestone, J. A. Doudna, A. Marson, Generation of knock-in primary human T cells using Cas9 ribonucleoproteins. Proc. Natl. Acad. Sci. U.S.A. 112, 10437–10442 (2015). doi:10.1073/pnas.1512503112pmid:26216948

23. M. Kurachi, R. A. Barnitz, N. Yosef, P. M. Odorizzi, M. A. DiIorio, M. E. Lemieux, K. Yates, J. Godec, M. G. Klatt, A. Regev, E. J. Wherry, W. N. Haining, The transcription factor BATF operates as an essential differentiation checkpoint in early effector CD8+ T cells. Nat. Immunol. 15, 373–383 (2014). doi:10.1038/ni.2834pmid:24584090

24. P. M. Odorizzi, K. E. Pauken, M. A. Paley, A. Sharpe, E. J. Wherry, Genetic absence of PD-1 promotes accumulation of terminally differentiated exhausted

CD8+ T cells. J. Exp. Med. 212, 1125–1137 (2015). doi:10.1084/jem.20142237pmid:26034050

25. B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10, R25 (2009). doi:10.1186/gb-2009-10-3-r25pmid:19261174

26. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550 (2014). doi:10.1186/s13059-014-0550-8pmid:25516281

27. P. Flicek, M. R. Amode, D. Barrell, K. Beal, S. Brent, Y. Chen, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gordon, M. Hendrix, T. Hourlier, N. Johnson, A. Kähäri, D. Keefe, S. Keenan, R. Kinsella, F. Kokocinski, E. Kulesha, P. Larsson, I. Longden, W. McLaren, B. Overduin, B. Pritchard, H. S. Riat, D. Rios, G. R. Ritchie, M. Ruffier, M. Schuster, D. Sobral, G. Spudich, Y. A. Tang, S. Trevanion, J. Vandrovcova, A. J. Vilella, S. White, S. P. Wilder, A. Zadissa, J. Zamora, B. L. Aken, E. Birney, F. Cunningham, I. Dunham, R. Durbin, X. M. M. Fernández-Suarez, J. Herrero, T. J. Hubbard, A. Parker, G. Proctor, J. Vogel, S. M. Searle, Ensembl 2011. Nucleic Acids Res. 39 (Database issue), D800–D806 (2011). doi:10.1093/nar/gkq1064pmid:21045057

28. O. L. Griffith, S. B. Montgomery, B. Bernier, B. Chu, K. Kasaian, S. Aerts, S. Mahony, M. C. Sleumer, M. Bilenky, M. Haeussler, M. Griffith, S. M. Gallo, B. Giardine, B. Hooghe, P. Van Loo, E. Blanco, A. Ticoll, S. Lithwick, E. Portales-Casamar, I. J. Donaldson, G. Robertson, C. Wadelius, P. De Bleser, D. Vlieghe, M. S. Halfon, W. Wasserman, R. Hardison, C. M. Bergman, S. J. Jones, Open Regulatory Annotation Consortium, ORegAnno: An open-access community-driven resource for regulatory annotation. Nucleic Acids Res. 36 (Database issue), D107–D113 (2008). doi:10.1093/nar/gkm967pmid:18006570

29. K. D. Pruitt, T. Tatusova, D. R. Maglott, NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 35 (Database issue), D61–D65 (2007). doi:10.1093/nar/gkl842pmid:17130148

30. X. Chen, H. Xu, P. Yuan, F. Fang, M. Huss, V. B. Vega, E. Wong, Y. L. Orlov, W. Zhang, J. Jiang, Y.-H. H. Loh, H. C. Yeo, Z. X. Yeo, V. Narang, K. R. Govindarajan, B. Leong, A. Shahab, Y. Ruan, G. Bourque, W.-K. K. Sung, N. D. Clarke, C.-L. L. Wei, H.-H. H. Ng, Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell 133, 1106–1117 (2008). doi:10.1016/j.cell.2008.04.043pmid:18555785

31. M. A. Hume, L. A. Barrera, S. S. Gisselbrecht, M. L. Bulyk, UniPROBE, update 2015: New tools and content for the online database of protein-binding microarray

data on protein-DNA interactions. Nucleic Acids Res. 43 (Database issue), D117–D122 (2015). doi:10.1093/nar/gku1045pmid:25378322

32. A. Mathelier, X. Zhao, A. W. Zhang, F. Parcy, R. Worsley-Hunt, D. J. Arenillas, S. Buchman, C.-Y. Chen, A. Chou, H. Ienasescu, J. Lim, C. Shyr, G. Tan, M. Zhou, B. Lenhard, A. Sandelin, W. W. Wasserman, JASPAR 2014: An extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Res. 42 (Database issue), D142–D147 (2014). doi:10.1093/nar/gkt997pmid:24194598

33. C. E. Grant, T. L. Bailey, W. S. Noble, FIMO: Scanning for occurrences of a given motif. Bioinformatics 27, 1017–1018 (2011). doi:10.1093/bioinformatics/btr064pmid:21330290

34. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842 (2010). doi:10.1093/bioinformatics/btq033pmid:20110278

35. J. Sunshine, M. Kim, J. M. Carlson, D. Heckerman, J. Czartoski, S. A. Migueles, J. Maenza, M. J. McElrath, J. I. Mullins, N. Frahm, Increased sequence coverage through combined targeting of variant and conserved epitopes correlates with control of HIV replication. J. Virol. 88, 1354–1365 (2014). doi:10.1128/JVI.02361-13pmid:24227851

36. M. R. Henn, C. L. Boutwell, P. Charlebois, N. J. Lennon, K. A. Power, A. R. Macalalad, A. M. Berlin, C. M. Malboeuf, E. M. Ryan, S. Gnerre, M. C. Zody, R. L. Erlich, L. M. Green, A. Berical, Y. Wang, M. Casali, H. Streeck, A. K. Bloom, T. Dudek, D. Tully, R. Newman, K. L. Axten, A. D. Gladden, L. Battis, M. Kemper, Q. Zeng, T. P. Shea, S. Gujja, C. Zedlack, O. Gasser, C. Brander, C. Hess, H. F. Günthard, Z. L. Brumme, C. J. Brumme, S. Bazner, J. Rychert, J. P. Tinsley, K. H. Mayer, E. Rosenberg, F. Pereyra, J. Z. Levin, S. K. Young, H. Jessen, M. Altfeld, B. W. Birren, B. D. Walker, T. M. Allen, Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. PLOS Pathog. 8, e1002529 (2012). doi:10.1371/journal.ppat.1002529pmid:22412369

37. S. M. Kaech, W. Cui, Transcriptional control of effector and memory CD8+ T cell differentiation. Nat. Rev. Immunol. 12, 749–761 (2012). doi:10.1038/nri3307pmid:23080391

38. M. Quigley, F. Pereyra, B. Nilsson, F. Porichis, C. Fonseca, Q. Eichbaum, B. Julg, J. L. Jesneck, K. Brosnahan, S. Imam, K. Russell, I. Toth, A. Piechocka-Trocha, D. Dolfi, J. Angelosanto, A. Crawford, H. Shin, D. S. Kwon, J. Zupkosky, L. Francisco, G. J. Freeman, E. J. Wherry, D. E. Kaufmann, B. D. Walker, B. Ebert, W. N. Haining, Transcriptional analysis of HIV-specific CD8+ T cells shows that PD-1 inhibits T cell function by upregulating BATF. Nat. Med. 16, 1147–1151 (2010).

doi:10.1038/nm.2232pmid:20890291

39. H. Shin, S. D. Blackburn, A. M. Intlekofer, C. Kao, J. M. Angelosanto, S. L. Reiner, E. J. Wherry, A role for the transcriptional repressor Blimp-1 in CD8(+) T cell exhaustion during chronic viral infection. Immunity 31, 309–320 (2009). doi:10.1016/j.immuni.2009.06.019pmid:19664943

40. G. J. Martinez, R. M. Pereira, T. Äijö, E. Y. Kim, F. Marangoni, M. E. Pipkin, S. Togher, V. Heissmeyer, Y. C. Zhang, S. Crotty, E. D. Lamperti, K. M. Ansel, T. R. Mempel, H. Lähdesmäki, P. G. Hogan, A. Rao, The transcription factor NFAT promotes exhaustion of activated $CD8^+$ T cells. Immunity 42, 265–278 (2015). doi:10.1016/j.immuni.2015.01.006pmid:25680272

41. M. Giordano, C. Henin, J. Maurizio, C. Imbratta, P. Bourdely, M. Buferne, L. Baitsch, L. Vanhille, M. H. Sieweke, D. E. Speiser, N. Auphan-Anezin, A.-M. M. Schmitt-Verhulst, G. Verdeil, Molecular profiling of CD8 T cells in autochthonous melanoma identifies Maf as driver of exhaustion. EMBO J. 34, 2042–2058 (2015). doi:10.15252/embj.201490786pmid:26139534

42. U. de Grazia, M. P. Felli, A. Vacca, A. R. Farina, M. Maroder, L. Cappabianca, D. Meco, M. Farina, I. Screpanti, L. Frati, A. Gulino, Positive and negative regulation of the composite octamer motif of the interleukin 2 enhancer by AP-1, Oct-2, and retinoic acid receptor. J. Exp. Med. 180, 1485–1497 (1994). doi:10.1084/jem.180.4.1485pmid:7931079

43. X. Feng, H. Wang, H. Takata, T. J. Day, J. Willen, H. Hu, Transcription factor Foxp1 exerts esse ntial cell-intrinsic regulation of the quiescence of naive T cells. Nat. Immunol. 12, 544–550 (2011). doi:10.1038/ni.2034pmid:21532575

3.7 Supplementary Information

Supplementary information, additional details on the methods and supporting figures are reproduced below. Additional tables can be found at: http://science.sciencemag.org/content/suppl/2016/11/04/science.aae0491.DC1?_ga=2.141521914.939072041.1533582113-83659740.1480874477

Materials and Methods

Mice

Wild type C57/BL/6N (CD45.2+) and congenic B6.SJL-Ptprca Pepcb/BoyJ (CD45.1+) mice were obtained from the US National Cancer Institute. C57BL/6N P14 mice (LCMV-specific T cell receptor transgenic) were bred in-house. Male mice were used at between 6 and 10 weeks of age. All animal experiments were done in accordance with the Animal Care and Use guidelines for the University of Pennsylvania.

Infection and isolation of mouse lymphocytes

LCMV strains were produced and titred as described (*23, 24*). Two cohorts of mice were infected by intraperitoneal injection of LCMV Armstrong (2×10^5 plaque-forming units) or intravenous injection of LCMV Clone 13 (4×10^6 plaque-forming units). All donor P14 cells were prepared from naive spleen using CD8 negative selection (Miltenyi Biotec). To generate effector, memory and exhausted $CD8^+$ T cells, approximately 1-2 x 10^3 P14 cells were transferred into recipient mice, followed by LCMV infection one day later. Target effector/memory/exhausted P14 cells were harvested only from recipient spleens at d8 or d27 p.i. After enrichment using CD45.2 FITC Ab and anti-FITC microbes (Miltenyi), cells were stained and sorted by Aria II (BD). Naïve and effector/memory/exhausted $CD8^+$ T (P14) cells were sorted by gating on $CD8^+$ $TCRVa2^+$ $CD62L^+$ $CD44^{lo}$ and $CD8^+$ $TCRVa2^+$ $CD45.2^+$ $CD45.1^-$ population, respectively. The following fluorochrome-conjugated antibodies from Biolegend were used for flow-cytometry: anti–TCR Vα2 (B20.1), anti-CD8α (53-6.7), anti-CD44 (IM7), anti-CD45.1 (A20), anti-CD45.2 (104), and anti-CD62L (MEL14).

ATAC-seq

We sorted 40-50,000 cells per biological replicate (Naïve, Acute d8, Acute d27, Chronic d8, Chronic d27 and EL4), which were then washed once in cold PBS and lysed in 50µL cold lysis buffer (10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM MgCl2, 0.1% IGEPAL CA-630). Lysed nuclei were incubated in Tn5 transposition reaction mix as described (*7*) and purified using MinElute Reaction Cleanup kit (Qiagen). ATAC-seq fragments from one set of replicates for Acute d8, Acute d27, Chronic d8 and Chronic d27, as well as both biological replicates for Naïve cells and EL4 cells, were size selected for fragments between 115 and 600 bp using Pippin Prep 2% Agarose Gel Cassettes and the Pippin Prep DNA Size Selection System (Sage Science). Post size-selection, ATAC libraries were amplified and Nextera sequencing primers ligated using Polymerase Chain Reaction (PCR). Finally, PCR primers were removed using Agencourt AMPure XP bead cleanup

(Beckman Coulter/Agencourt) and library quality was verified using a Tapestation machine. High quality 'multiplexed' DNA libraries were sequenced on the Illumina HiSeq2000.

Quality control of ATAC-Seq data and alignment to the genome

We used the FASTQC pipeline (Babraham Bioinformatics) on our reads, and aligned them to the reference genome (mm9) with bowtie version 1.1.1 (*25*), retaining only reads that mapped to a unique position in the genome ["–m 1"]. We marked duplicate reads in the bam files using PICARD and checked for contamination of primer sequences.

Identification of peaks in ATAC-Seq data

After alignment of the reads to the reference genome, reads aligned to the positive strand were moved +4 bp, and reads aligning the negative strand were moved -5bp (*7*). For each of our five replicate pairs, we merged the two replicates using samtools, and then called peaks using MACS2 version 2.1.0, setting the FDR to 0.001 with the "nomodel" option and default mouse genome size. To generate our final universe of peaks, we took the union of all of these peaks while merging overlapping regions.

Identification of differential activity in peaks

We extracted the shifted ATAC-Seq cut sites from our data and counted the number that fell into each peak region. We then used DESeq2 (*26*) to identify the differential abundance of cut sites between the peaks and performed all pairwise comparisons of the five conditions and considered those with a FDR below 0.05 to be differential. The data presented in the ATAC-seq heat maps (Figures 2A, 4C), scatterplot (Fig S1C), and PCA (Fig S1D) is the number of cutsites in each peak region scaled according to the DESeq2 scaling factor.

Enrichment of genomic features in ATAC-Seq peaks

For each sample, we investigated the set of peaks calculated by MACS2 for enrichment of genomic features relative to their coverage in the mouse genome. For each genomic feature, we calculated the fold enrichment as [Overlap of Genomic Feature with ATAC-Seq Peak Regions / Total Size of Peak Regions] / [Overlap of Genomic Feature with Mouse Genome / Effective Size of Mouse Genome], with all sizes expressed in base pairs.

We set the effective size of the mouse genome to be 2,716,965,481 bp. *P*-values were calculated using a binomial test where n= Number of Peaks, x = Number of Peaks Overlapped by Annotation, p = (Overlap with Mouse Genome / Effective Size of the Mouse Genome).

The genomic features we used include: (i) regulatory features annotations from the Ensemble database (*27*), (ii) regulatory features found by the ORegAnno database (*28*), (iii) conserved regions annotated by the multiz30way algorithm, here we consider regions with multiz30way score>0.7, (iv) repeat regions annotated by RepeatMasker

(http://www.repeatmasker.org), (v) putative promoter regions - taking 10 kb upstream and 1 kb downstream of transcripts annotated in RefSeq (*29*), (vi) gene body annotations in RefSeq; (vii) 3' proximal regions (taking 1kb upstream and 5kb upstream to 3' end); (viii) regions enriched in binding of BATF, IRF4, c-Jun, JunD, and JunB, along with histone marks H3K4me1, H3K4me3, H3K27ac, H3K36me3, and H3K27me3 (Polycomb-repression), and additional annotations Poised Enhancer, Active Enhancer, Bivalent Promoter, Developmental Enhancer, and Init. Promoter in CD8+ T Cells (*23*). We also added the peaks we obtained in this study for two replicates of EL4 cells. These results are summarized in Table S1 and Figure S1G.

Cluster and gene ontology analysis

K-means clustering was applied using GENE-E (Broad Institute) to differential ChAR signal intensity across all five cell states. Optimum number of clusters was determined using the gap statistic in MATLAB. Association of ChARs to neighboring genes and enrichment of Gene Ontology (GO) terms within modules (using the entire genome as a background set) was determined using GREAT with default settings (http://bejerano.stanford.edu/great/public/html/). All significant GO terms are enumerated in table S2, while selected GO terms are presented as a heatmap in Fig 2D.

CRISPR design

sgRNA-specifying oligo sequences were chosen to maximize on-target efficiency based on publicly available on-line tools (http://www.broadinstitute.org/rnai/public/analysis-tools/sgrna-design-v1). Single stranded sgRNA oligos (Integrated DNA Technologies) were annealed and ligated into PXPR003 as previously described (*14*). SgRNA plasmids were introduced into cell lines using lentiviral packaging or nucleofection. For lentivirus production, 293T cells were seeded in DMEM with 10% (vol/vol) FBS. Cells were transfected with sgRNA plasmids and the packaging plasmids Δ8.9 and VSV-g using TurboFect (ThermoFisher). Viral supernatants were collected 72 hours later. For nucleofection, EL4 cells were mixed with 100ul of Nucleofector solution and 2ug of sgRNA plasmid. Cells were then transfected using Nucleofector Kit L (Lonza) and the C-09 program on the Amaxa Nucleofector I machine.

Cell culture

EL4 cells (ATCC) were cultured in RPMI with 10% (vol/vol) FBS, HEPES, Penicillin/Streptomycin and β-mercaptoethanol (R10 media). Cas9 expressing cell lines were produced using PLX304 plasmid lentivirally packaged as described above. Cas9 expressing cells were then transduced or transfected with sgRNA plasmids, selected for plasmid presence using puromycin, and used for all subsequent experiments 3-5 days post-selection. Single cell clones of bulk EL4s were made using limiting dilution.

Reporter assays

DNA sequences corresponding to the 781 bp core of the -23.8kb PD-1 enhancer (chr1: 95971118 - 95971899, mm9), and the CMV promoter were cloned upstream of a TATA

box minimal promoter driving GFP expression within a lentiviral construct. Plasmids were lentivirally packaged as described above. Viral supernatant was concentrated 100x by ultracentrifugation for 2 hours at 20,000xg.

$CD8^+$ T cells were isolated from naive spleens using CD8 negative selection MACS kit (Miltenyi Biotec) and cultured in R10 medium. Naive cells were stimulated for 24 hours with plate-bound anti-CD3 (4 μg/ml; 2C11; BD Pharmingen) and anti-CD28 (4 μg/ml; 37.51; BD Pharmingen) in the presence of recombinant human IL-2 (100 U/ml, R&D Systems) for *in vitro* activation prior to transduction. *In vitro* activated T cells and EL4 cells were transduced with concentrated virus by incubation with polybrene (5ug/ml) and spin infection (2000 RPM for 45 min at 37 °C). GFP expression was measured by flow cytometry 48-72 hours post transduction.

Enhancer deletion

The region -23.8kb from the *Pdcd1* TSS was deleted using a pair of sgRNA targeting sequences flanking the region. EL4 cells transfected with -23.8kb-targeting sgRNAs or negative control sgRNAs (fig. S5A) were sorted on the basis of differing levels of PD-1 expression and genomic PCR was used to detect deletion in sorted EL4s. Genomic DNA was isolated from bulk and clonal cell lines using 50ul QuickExtract DNA solution and incubated according to manufacturer instructions. PCR was performed on genomic DNA from each sample using the deletion screening primer pairs (fig. S5B), and run on a 1% agarose gel with EtBr. For clonal lines, biallelic deletion clones were defined as having PCR amplification of only the deletion band, while non-deleter clones amplified only the WT band. Deletion was verified by submitting the PCR amplification products for Sanger sequencing.

RT-qPCR

Clonal cells lines as well as control sgRNA receiving bulk populations were lysed in RLT and RNA was extracted using RNeasy Plus Mini kit (Qiagen). cDNA was generated using ImProm-II Reverse Transcriptase (Promega), and analyzed on a Viia 7 Quantitative PCR instrument using Taqman probes.

***In situ* saturation mutagenesis**

All possible tiling sgRNAs as constrained by Cas9 –NGG protospacer adjacent motif presence within all nine PD-1 enhancers and all PD-1 exons were designed using publicly available online tools (http://portals.broadinstitute.org/gpp/public/analysis-tools/sgrna-design). Non-targeting sgRNAs were designed by the Genomics Perturbation Platform at the Broad Institute. All sgRNA oligos were synthesized and cloned into the lentiGuide-Puro vector backbone as previously described (*14, 15*). Plasmid libraries were lentivirally packaged as described above, and Cas9-expressing EL4s were transduced by incubation with polybrene (5ug/ml) and spin infection (2000 RPM for 45 min at 37 °C). Control transductions were done to ensure 1000x library coverage with a multiplicity of infection of approximately 0.3. Transduced cells were selected with puromycin for 72 hours, and a sample was taken post-selection as the starting d0 representation of the

library. Cells were cultured throughout the experiment to ensure 1000x coverage of the library at all times.

Library-transduced cells were stained with anti-PD-1 fluorescent antibodies and sorted into PD-1 high and PD-low populations (fig. S4C). Replicate sorts were done between d20-d50 post-selection, and genomic DNA was isolated from sorted populations using the Dneasy Blood and Tissue Kit (Qiagen). Bulk cells from d0 post-selection, input plasmid pools as well as sorted populations were sequenced as previously described (*14, 15*). SgRNA abundance was quantified within each sample by normalizing reads to library sequencing depth. SgRNA enrichment in PD-1 high vs. PD-1 low populations were calculated as follows: 1) sgRNA abundance in each sample was log-transformed; 2) sgRNA abundance at d0 was subtracted from its abundance in each sample to account for initial variation in sgRNA representation; 3) any sgRNA that was not detected in at least one sorted sample was excluded from further analysis as a cellular dropout; 4) for each sorted sample (high and low analyzed separately), the abundance of enhancer- and exon-targeting sgRNAs were normalized calculating z-scores relative to the distribution of the control guides; 5) the normalized score for each sgRNA in the PD-1 low compartment was subtracted from the score in the PD-high compartment, to generate a normalized enrichment score for each sgRNA for each replicate; and 6) the enrichment score for each sgRNA was averaged across replicates. Enhancer- and exon-targeting sgRNA cleavage sites were mapped to the mouse genome (mm9), while control non-targeting sgRNAs were pseudo-mapped with 5 bp spacing.

Validation of individual sgRNAs

32 sgRNAs that were enriched in the PD-1 low fraction within the -23.8kb enhancer were chosen for individual validation. Oligos representing each sgRNA (IDT) were cloned individually into Cas9 and sgRNA dual delivery construct. Plasmids were then nucleofected into EL4 cells as previously described to produce 32 isogenic cell lines. Each cell line was selected 18-36 hours post-nucleofection in puromycin, and cultured in R10 medium. Cells were stained for PD-1 expression or the corresponding isotype control. Geometric mean fluorescence intensity (MFI) of PD-1 expression in each cell line was normalized to the MFI of staining with isotype control.

Identifying motifs in genome

We used the PWM databases from Jolma 2013 , Chen 2008 (*30*), UniPROBE (mouse) (*31*), and JASPAR 2014 (core vertebrates) (*32*) available in the list of databases at http://meme-suite.org/db/motifs. We then called motifs in version 9 of the mouse genome using FIMO (*33*) and a threshold of $1x10^{-5}$. We began with a total of 1,446 motifs and removed any that had zero hits, or more than 1.5 million hits.

We then added evolutionary conservation and distance to TSS for each motif. We took the mean evolutionary conservation found in each motif, and used bedtools (*34*) “closest” function to find the nearest TSS for each motif. A file of TSS was created from the UCSC genome browser.

Inferring transcription factor footprints using Centipede

We then used the program Centipede to infer the posterior probability of transcription factor binding at each motif instance. We passed the evolutionary conservation score, distance to TSS, and PWM score for each motif, along with the number of ATAC-Seq cuts found in every position in a 220 bp window surrounding the motif, on both strands separately.

As an additional filtering step, we ran a separate instance of Centipede without conservation and regressed its posterior probability on evolutionary conservation, distance to TSS, and PWM score and removed those that did not have a strong, positive association with evolutionary conservation in at least one sample ($P < 0.05$ for one-sided Z value test of coefficient). To evaluate the accuracy of these predictions, we compared the results to available ChIP-seq data (Fig S6C-D).

Identifying association of TF with chronic vs. acute infection

To test whether the binding locations predicted for a given TF are associated with T cell exhaustion, we conducted hypergeometric tests comparing the chronic vs. acute infection in each time point (Table S3). A TF was marked as differentially bound in a given condition if the TF was found more often than expected by chance in peaks that were active in that condition versus the other condition. To identify TF associated with the transition form the naïve state, repeated a similar analysis comparing naïve vs. day 8 acute infection (Fig 7A).

Ranking of TFs

The normalized mRNA expression of each TF was obtained from previously published datasets (*3, 8*), and the absolute value of log fold change between Acute and Chronic infection was calculated at d7 and day 30. TFs were then associated with a hypergeometric *P*-value quantifying differential binding at d8 and d27 (as above; taking the most significant score across all motifs that correspond to that TF). Each TF was then ranked from most to least differential for these four annotations (taking the average rank), representing changes in gene expression and TF binding at d7-8 and d27-30. The final ranking for each TF was determined as the average of the four individual scores. The information in Tables 5-6 is provided for the top 100 TFs, plus addition of six TFs with known relevance to $CD8^+$ T cell biology (JUND, RUNX3, JUNB, STAT5A-STAT5B, FOXP1, TCF7).

Analysis of TF binding to regions

Binding at a transcription factor motif within a region was inferred from Centipede using a posterior probability cutoff of 0.90 or greater. The binding genome-wide was inferred for each of 106 transcription factors motifs, which were determined by their ranking (top 100, as above; tables S3-6) and known relevance in $CD8^+$ T cell biology (JUND, RUNX3, JUNB, STAT5A-STAT5B, FOXP1, TCF7). Those 106 TFs were further filtered to 50

based on their differential binding and known role in $CD8^+$ T cell transcriptional regulation.

We then looked at the binding at each TF motif instance of the 50 TFs in the acute and chronic state. Each binding event was classified as being present in acute only, chronic only, or both. This comparison between motif binding in acute or chronic infection was done at both d8 and d27. Thus at d8 and d27, we were able to calculate the percentage of binding events per TF motif that was unique to chronic infection, acute infection or shared. This partiality of TF binding in either acute or chronic infection only, as opposed to shared binding, was quantified as the percentage of all unique binding events seen in chronic infection at each time-point. Furthermore, each region containing at least one bound motif was associated with a target gene using GREAT. All target genes could then be defined as being upregulated, downregulated or unchanged by mRNA expression between acute and chronic infection at d7 and d30 (*8*). Thus, each binding event between a TF motif and a near target gene (defined using a more stringent probability cutoff of >0.95) was classified into those that targeted a gene upregulated in acute infection, upregulated in chronic infection or unchanged (table S3-6).

Subjects

Four HIV-infected participants in the Seattle Natural Progression cohort were selected for this study (*35*). Progressors were defined as having median viral loads >10,000 RNA copies/ml in the last year. All subjects were studied in chronic infection and were antiretroviral therapy naïve. 40-60 million PBMCs were obtained from each subject. The relevant institutional review boards approved all human subject protocols, and all subjects provided written informed consent before enrollment.

A baseline sample was obtained for a single HCV-infected patient enrolled on a trial to evaluate the effect of successful antiviral therapy on innate and adaptive immune responses for genotype 1a hepatitis C virus infection (NCT02476617). The HCV subject was chronically infected with an HCV viral load of 828,000 IU/ml. Based on the expressed HLA class I alleles, we screened for virus-specific CD8 T cell responses using HLA class I multimers. We identified responses targeting HCV A*02:01 C63B CINGVCWTV, HCV A*24:01 174D VIAPAVQTNW and influenza A*02:01 MP GILGFVFTL. Multimer-binding CD8 T cells were isolated by FACS from PBMCs obtained through leukapheresis. The patient had consented to this Institutional Review Board approved study.

PBMCs were obtained via density centrifugation from a normal volunteer and screened for a CMV $tetramer^+$ T cell response. The relevant institutional review board approved the human subject protocol, and all subjects provided written informed consent before enrollment.

Isolation of human lymphocyte populations for ATAC-seq

PBMCs were thawed rapidly in warm RPMI supplemented in 10% FBS. $CD8^+$ T cells were enriched using MACS CD8 negative selection kit (Miltenyi Biotec). After enrichment, cells were stained and sorted on a FACSAria cell sorter (BD Biosciences); see fig. S7A for cell sorting strategy. The following fluorochrome-conjugated antibodies from Biolegend were used for flow-cytometry: anti–CD45RA (HI100), anti-CCR7 (G043H7), anti-2B4 (C1.7), anti-CD3 (OKT3), anti-CD39 (A1), anti-PD-1 (EH12.2H7), and anti-CD8a (SK1). The following fluorochrome-conjugated multimers were used for flow-cytometry: HIV A*03:01 RLRPGGKKK, HIV A*02:01 SLYNTVATL, HIV B*08:01 GEIYKRWII, CMV A*02:01 pp65-NLV NLVPMVATV, HCV A*02:01 C63B CINGVCWTV, HCV A*24:01 174D VIAPAVQTNW, and influenza A*02:01 MP GILGFVFTL. A maximum of 70,000 cells were sorted PBS supplemented with 10% FBS and ATAC-seq libraries were generated as previously described. Due to limiting starting material, ATAC libraries were not size-selected prior to multiplexed sequencing on the Illumina HiSeq2000. Post-sequencing quality control, alignment of human ATAC reads to the human genome (hg19), MACS2 peak calling, and DESeq2 peak normalization was done as previously described.

Mapping peaks to human genome and SNP analysis

Orthologous mouse ChARs (mm10) were mapped to the human genome (hg19) as described (*18*). Since the mapping algorithm requires input regions in mm10, the UCSC liftover tool was applied to ChARs to transfer them onto mm10 from mm9.

GWAS SNPs in the NHGRI catalog (https://www.ebi.ac.uk/gwas/) annotated as “immune system disease” was defined as being immune-related. Hypergeometric tests were performed quantifying the overlap of mapped enhancers with all GWAS SNPS, immune-related SNPS as well as PICS SNPs.

Comparative analysis of mouse and human data

All mouse peaks identified across the 5 cell states (Naïve, Acute d8, Acute d27, Chronic d8, Chronic d27) were partitioned into 3 categories based on MACS2-called peaks: peaks identified uniquely in Naïve cells relative to Acute and Chronic d27 (mouse Naïve only), peaks identified uniquely in Acute d27 cells relative to Naive and Chronic d27 (mouse Memory only), and peaks identified uniquely in Chronic d27 cells relative to Naive and Acute d27 (mouse Exhaustion only). All mouse peaks not classified into one of those 3 categories were excluded from further analysis. Orthologous mouse peaks within the three categories (Naïve, Memory, Exhaustion) were mapped to the human genome and filtered for overlap with at least one MACS2 human peak called independently across all human samples. Then, for each human sample, average chromatin accessibility was calculated separately for the 3 categories of mouse orthologous peaks (Naïve, Memory, Exhaustion). Average chromatin accessibility within each category was normalized to account for inherent differences in chromatin accessibility across the 3 classes of peaks.

Amplification and deep sequencing of HCV epitopes

For HCV epitopes 174D and C63B, amplicons surrounding these epitopes were generated using the following conditions. The reaction consisted of 2x First Strand

Buffer, sense (A2F: AAC GTT GCG ATC TGG AAG AC or A3F: GCT CTC ATG ACC GGC TTT AC) and antisense primers (A2R: GGA AGC GTG GTT GTC TCA AT or A3R: AGA GAT CTC CCG CTC ATC CT) at 0.4μM, and a Superscript III RT/Platinum *Taq* Mix (Invitrogen), with the following conditions: cDNA synthesis for 30 minutes at 50°C, followed by heat denaturation at 94°C for 2 minutes, the PCR amplification conditions were 40 x (94°C, 15 s; 55°C, 30 s; 68°C 180 s), with a final extension at 68°C for 5 minutes. PCR amplicons were fragmented and barcoded using NexteraXT DNA Library Prep Kit, as per manufacturer's protocol. Samples were pooled and sequenced on an Illumina MiSeq platform, using a 2 x 250 bp V2 reagent kit. Paired-end reads obtained from Illumina MiSeq were assembled into a HCV consensus sequence using the VICUNA *de novo* assembler software and finished with V-FAT v1.0 (*36*). Reads were mapped back to this consensus using Mosaik v2.1.73, and intra-host variants called by V-Phaser v2.0 (*36*). All reads have been deposited to the NCBI Sequence Read Archive under accession number SRR3951347.

Supplementary Tables

Table S1. Chromatin accessible regions (ChARs). This table lists the 43,171 Chromatin Accessible Regions identified in $CD8^+$ T Cells. Each row represents a ChAR and the columns contain the number of ATAC-Seq cuts, *P*-values and FDRs for differential abundance of cuts (calculated by DESeq2), overlap with transcription factor binding sites based on ChIP-seq, overlap with annotations for the mouse genome, and information on nearby genes.

Table S2. Enrichment of GO terms in ChAR modules. This table lists all significantly enriched GO terms within the five, state-specific ChAR modules. Each row represents an individual GO term, and columns represent the significance of the enrichment (hypergeometric test, GREAT online tool) in each of 5 modules.

Table S3. Enrichment of transcription factor binding in $CD8^+$ T cell states in acute vs. chronic infection. This table examines the footprints inferred for each DNA binding motif (Centipede posterior binding score was over 0.90) and measures if the set of ATAC-seq peaks that coincides with these footprints is enriched for peaks upregulated in the Acute or Chronic condition.. Each row provides the results for a particular motif and the comparison "ChronicD8 vs. AcuteD8" or "ChronicD27 vs. AcuteD27". Using the results from DESeq2, we partition the peaks that coincides with the footprint as upregulated in the Acute condition, upregulated in the Chronic condition, or neither. Fold Enrichment in the Acute case is defined as [(Acute Peaks with footprint / Total Acute Peaks) / (Total Peaks with footprint / All Peaks)], and the *P*-value is derived from the hypergeometric distribution as one minus the probability of the footprint coinciding with at least as many of the Acute peaks found in the data by chance. The Fold Enrichment and *P*-value are calculated similarly for the Chronic condition.

Table S4. Differential gene expression in CD8$^+$ T cells from Acute and Chronic infection. This table provides changes in mRNA expression between Acute and Chronic conditions. Each row represents a single gene in the mouse genome, and the columns provide the log fold change in mRNA expression from CD8$^+$ T cells between Acute vs. Chronic infection at d30 and d7 based on microarray data from (*3, 8*)

Table S5. Ranking of transcription factors. This table lists the ranking of the TFs based on their differential expression (using microarray data in Table S4) and inferred differential binding (using footprint analysis from Table S3). Each row represents a transcription factor, and the columns provide the list of motifs used for ATAC-Seq footprinting, PD-1 enhancer, known gene, rank of the TF by absolute log fold change of expression for Acute vs. Chronic d30 and Acute vs. Chronic d8, and rank of the TF by hypergeometric test for differential binding in the comparisons Acute d8 vs. Chronic d8 and Acute d27 vs. Chronic d27. The overall rank is taken as the mean value of the four comparisons in each row.

Table S6. Region level binding of top TFs. This table shows inferred binding of top ranked TFs in each ChAR in each cell state. Each row represents one ChAR and each column represents the activity of a given TF in a given cell state. The value in each cell corresponds to the Centipede posterior probability of the binding of a specific TF in a specific cell state in a specific region, taking the maximum over all the corresponding motifs.

Table S7. Patient characteristics and viral load information. This table summarizes the characteristics of all subjects who were profiled in this study, including gender, HLA allele and CD4$^+$ T cell to CD8$^+$ T cell ratio (when available). Viral load for HIV and HCV patients are reported as plasma RNA IU/ml.

Supplementary Figure Legends

Figure S1. Characteristics of chromatin accessible regions from mouse CD8$^+$ T cells. **(A)** Schematic diagram of the experiment. **(B)** Number of chromatin accessible regions per condition, partitioned into those overlapping transcriptional start sites (TSS), exons, introns and intergenic areas as indicated. **(C)** Scatterplot showing correlation in peak intensity between biological replicates for Naïve, Acute d8, Acute d27, Chronic d8 and Chronic d27 states. (**D**) Principal component analysis of biological replicates for Naïve, Acute d8, Acute d27, Chronic d8 and Chronic d27 states across all 43,171 ChARs. **(E)** ATAC-seq fragment sizes from all five cell states. **(F)** Distance to nearest TSS for all intergenic chromatin accessible regions (as percentage of total), for all five cell states. **(G)** Fold enrichment of regions annotated for evolutionary conservation, regulatory region status and histone marks within ATAC peaks, for all five cell states. **(H)** Combined ATAC signal across all TSSs (black), H3K27ac peaks (dark grey) and H3K27me3 (light grey). Histone mark peaks were determined from (*23*).

Figure S2. Differential chromatin accessible regions in CD8$^+$ T cells from acute and chronic infection. (A) Fraction of constant and changing chromatin accessible regions. **(B)** Fraction of the genome covered by chromatin accessible regions over time, in naïve (grey symbol), CD8$^+$ T cells and those following Acute (white symbols) or Chronic (black symbols) infection. **(C)** Number of shared and differential chromatin accessible regions in CD8$^+$ T cells from acute and chronic infection at day 8 and day 27 (FDR < 0.05 for differential) for combinations indicated by the grey/white symbols below. **(D)** Distance to nearest TSS for all intergenic regions that are non-differential (light grey) and differential (dark grey).

Figure S3. Comparison of differential regions between states. (A) Graph showing the gap statistic using given number of K-means clusters, applied to the signal intensity of all differential regions across the five cell states. **(B)** Correlation between average gene expression and average chromatin accessibility for each module in each cell state. Letter symbols correspond to modules as indicated in Figure 4(A) and (B) in the main text.

Figure S4. CRISPR/Cas9-mediated enhancer editing. (A) ATAC-seq tracks from CD8$^+$ T cells, EL4 cell line and hematopoietic lineages indicated (*12*). Arrows indicate individual ChARs. **(B)** GFP reporter expression in EL4s (upper) and activated CD8$^+$ T cells (lower) transduced with a negative control minimal promoter (left), CMV promoter construct (middle) and *Pdcd1* -23.8 kb enhancer construct (right). Averages of 3-4 replicates are shown in the bar graphs. **(C)** Schematic diagram of genome editing strategy. **(D)** CRISPR PD-1 targeting and non-targeting sgRNAs used in EL4 cells. **(E)** PCR primers used for screening genomic deletion and for Sanger sequencing. **(F)** Histogram of PD-1 expression in EL4 cells transfected with control (light grey) or double-cut sgRNAs (red) targeting the -23.8kb region. **(G)** Gel showing PCR screening results using genomic DNA from control sgRNA, single enhancer sgRNA, and double enhancer sgRNA receiving EL4 cells. **(H)** Gel showing PCR screening results using genomic DNA from 45 EL4 single cell clones receiving double enhancer targeting sgRNAs. **(I)** Expression of PD-1 in EL4 single-cell clones that are WT (white bars) or Del (red) for the -23.8kb enhancer ($P < 0.0002$, Mann-Whitney rank-sum test for PD-1 expression). **(J)** Sequence at the enhancer for representative EL4 single cell clones. Clone numbers correspond to (E), arrows indicate expected cut sites from sgRNAs. **(K)** Relative mRNA expression in WT or Del clones ($P < 0.005$, T-test).

Figure S5. *In situ* saturation mutagenesis of the PD-1 enhancers. (A) Composition of sgRNAs within pooled screening library. **(B)** Distribution of gaps between adjacent sgRNA cut sites. **(C)** PD-1 distribution (left) in pooled sgRNA transduced EL4s (red) compared to control-transduced EL4s (grey). Gates indicate PD-1 high and low sorted fractions. Post-sort distribution of PD-1 in high and low sorted populations (right). **(D)** Correlation of sgRNA representation in plasmid pool vs. in post-transduction EL4s. **(E)** SgRNA enrichment scores within PD-1 high and low populations across 3 replicates. **(F)** Normalized enrichment of all sgRNAs (grey symbols) within PD-1 high and low populations in -23.8kb enhancer, with 10bp running average of sgRNA enrichment

(black). **(G)** Sequences for sgRNAs enriched in the PD-1 low fraction within the -23.8kb enhancer, which were chosen for individual validation. Relative positions from the start of the -23.8kb enhancer are shown on left. **(H)** Correlation between normalized enrichment score in main Figure 3(D) and PD-1 MFI of individually transfected EL4 cells, for top-scoring sgRNAs (PD-1 high:low ratio >1SD below mean) in the -23.8kb enhancer.

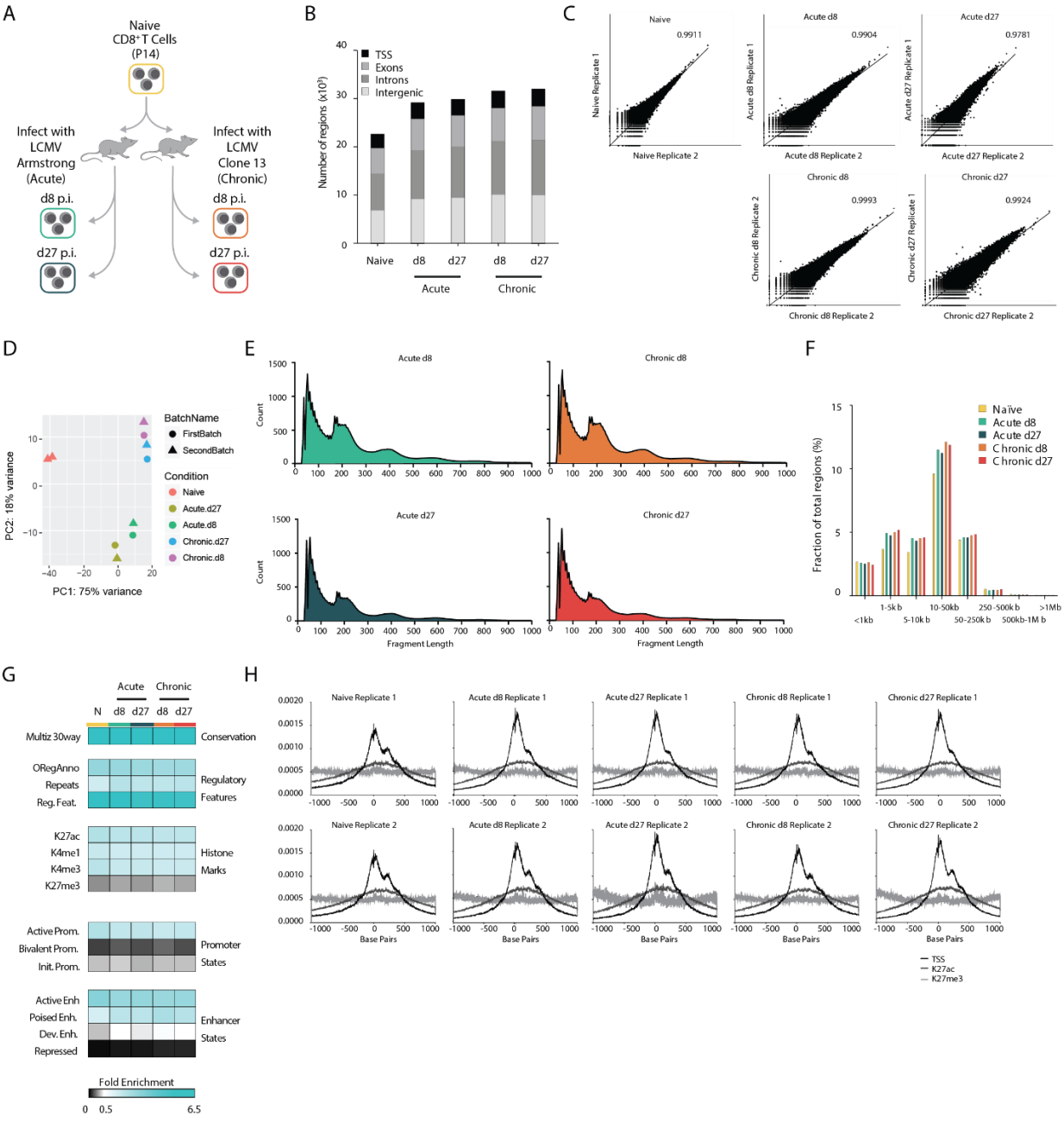
Figure S6. Evaluation of transcription factor footprinting with ChIP-Seq data. (A) IGV tracks showing ATAC-seq peaks for Naïve, Acute d8, Acute d27, Chronic d8 and Chronic d27 cell states, as well as c-Jun, Batf and IRF4 ChIP tracks from *in vitro* activated $CD8^+$ T cells, at *Gzmb* locus (*23*). **(B)** ATAC-seq profiles were analyzed to infer TF binding based on an attenuation of transposase cut sites over known TF motifs i.e. TF footprinting (*17*). Representative transcription factor (TF) footprints TFs are indicated. **(C)** TF footprinting was validated by plotting correlation of normalized ChIP-Seq reads and Centipede posterior probability for BATF and IRF4 across the full genome. **(D)** ROC plots of varying cutoff for Centipede posterior probability used to call presence/absence of IRF4 and BATF binding (ChIP-Seq peaks used as standard for identifying TF binding). These plots indicate high levels of sensitivity and specificity (AUC = 0.77 and 0.84).

Figure S7. Transcription factor footprinting in chromatin accessible regions. (A) Log fold enrichment of predicted TF footprints in acute d8 v. naive $CD8^+$ T cells (X-axis) are plotted against the corresponding P-value (hypergeometric). This analysis recovered motifs associated with known regulators of effector function including the AP-1 family TFs Batf, Jun, and Fos and the naïve T cell regulator Lef1 (*23, 37*). **(B)** Motifs for Sox3, RAR and TBX21 with footprints in the -23.8kb PD-1 enhancer **(C)** Network diagram showing inferred binding at TFs motifs (round nodes) in regions adjacent to genes (square nodes). Genes are partitioned by mRNA expression into those upregulated (left), downregulated (right) or unchanged (center) between acute and chronic infection, at day 7 (top) and day 27 (bottom) post infection. TF nodes are colored to indicate partiality of state-specific binding in chronic (red) or acute (green) infection. Edges colored to indicate relative fraction of state-specific binding events to the 3 classes of genes (upregulated in acute infection, chronic infection or unchanged). Edges denoting binding-events that are identified preferentially in chronic infection are in red, while those seen preferentially in acute infection are shown in green.

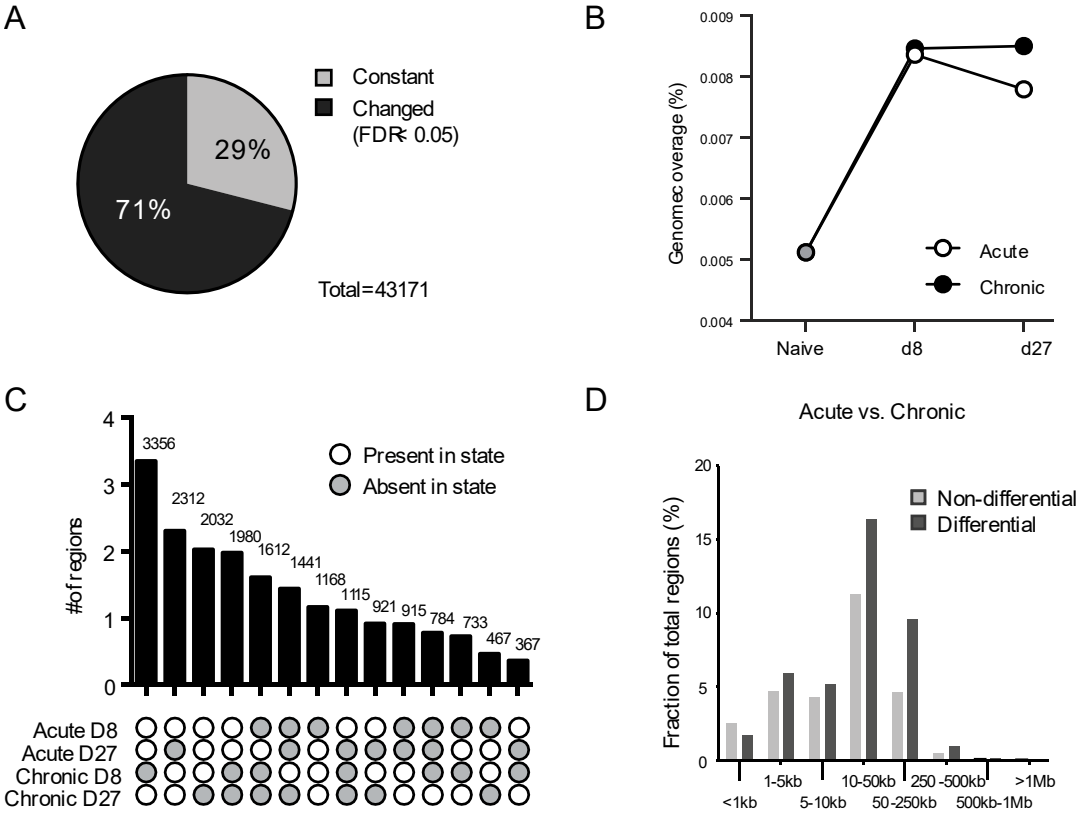
Several features suggest that binding at these TF motifs is likely to play a role in regulating T cell exhaustion. First, bound motifs are observed more frequently at genes differentially expressed between $CD8^+$ T cells in chronic vs. acute infection at d27 p.i. than in genes that were not differentially expressed (hypergeometric $P = 9.85 \times 10^{-3}$, differential gene expression *FDR* < 0.05). Second, many bound motifs recovered by our analysis correspond to TFs with known roles in exhaustion, including T-bet (*9*), Eomes (*9*), Batf (*38*), Blimp1 (*39*), Nfatc1 (*40*), Nr4a2 (*41*), Mafb (*41*). Third, other differential TF footprints suggest plausible, novel candidate regulators of T cell exhaustion, such as: Pou2f2 (also known as Oct-2) which interacts with AP-1 family TFs (*42*), and Foxp1, which regulates quiescence in naïve $CD8^+$ T cells (*43*).

Figure S8. Characteristics of human CD8$^+$ T cell chromatin accessible regions and mouse orthologous regions. (A) Cell sorting strategy for tetramer$^+$, naïve and effector memory populations on a representative human sample. **(B)** Fold enrichment of regions annotated for published histone marks in human primary T cells, within ATAC peaks for all human samples. **(C)** Percentage of all peaks entirely/partially mapped to human genome (hg19), for naïve, acute d8, acute d27, chronic d8 and chronic d27 states. **(D)** Enrichment of PICS SNPs in mouse orthologous ChARs. Dotted line illustrates $P = 0.05$ (hypergeometric test). **(E)** Enrichment of all NHGRI GWAS SNPs and **(F**) of immune system disease related SNPs. Dotted line illustrates $P = 0.05$ (hypergeometric test). **(G)** Heatmap showing fold enrichment of autoimmune diseases SNPs in orthologous regulatory regions relative to randomly selected regions; columns are hierarchically clustered. **(H)** Schematic diagram of mouse chromatin accessible region mapping to human genome. Regions of chromatin accessibility are defined in the mouse genome (left panels). Regions orthologous to the mouse regions are mapped to the human genome, and chromatin accessibility of the orthologous regions assessed in human T cell data (right panels).

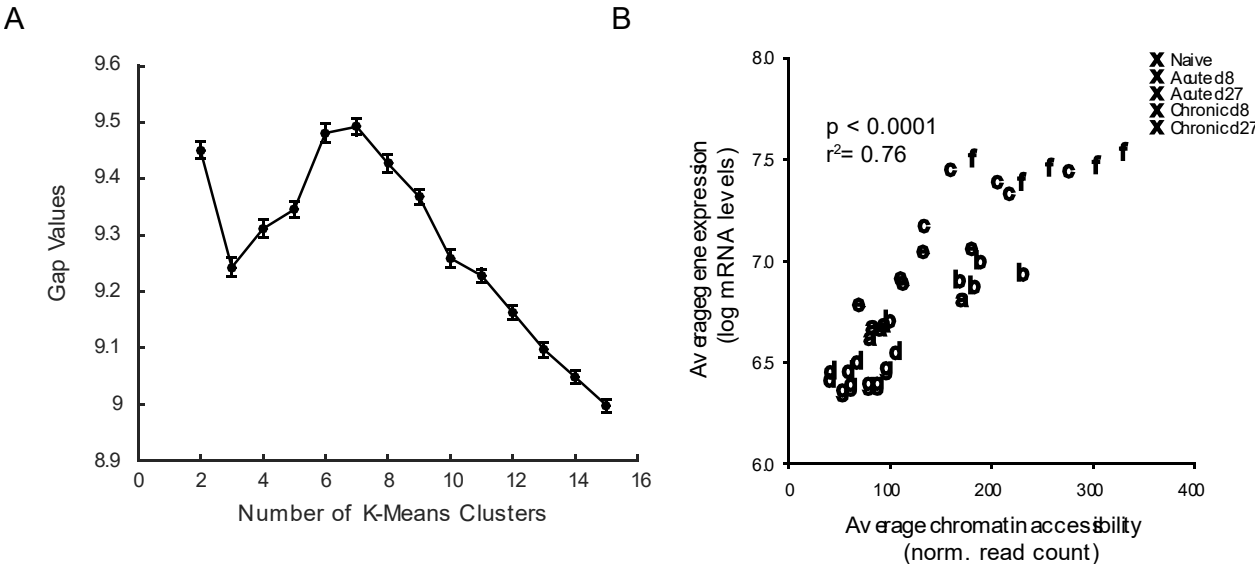
Supplementary Figure 1



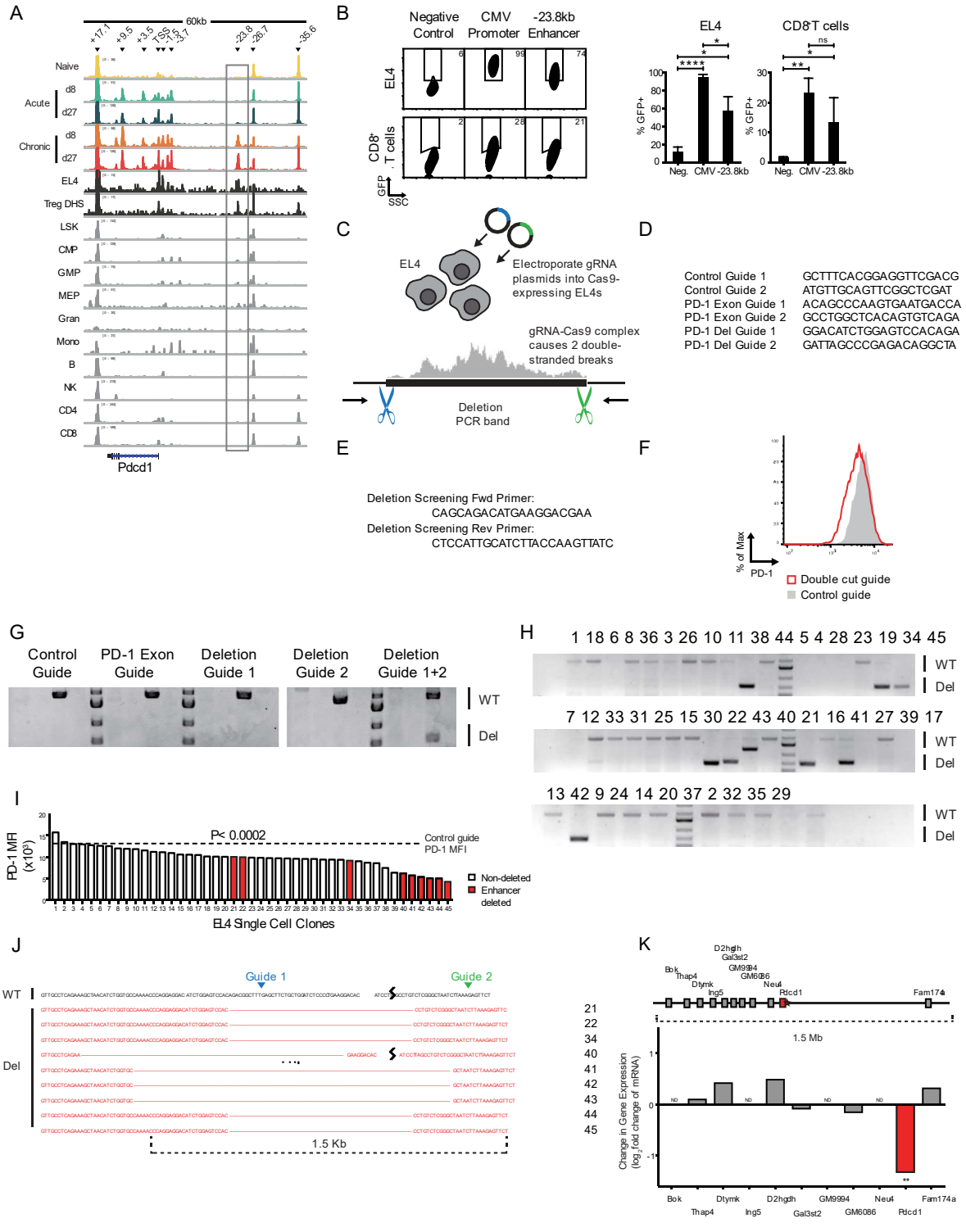
Supplementary Figure 2



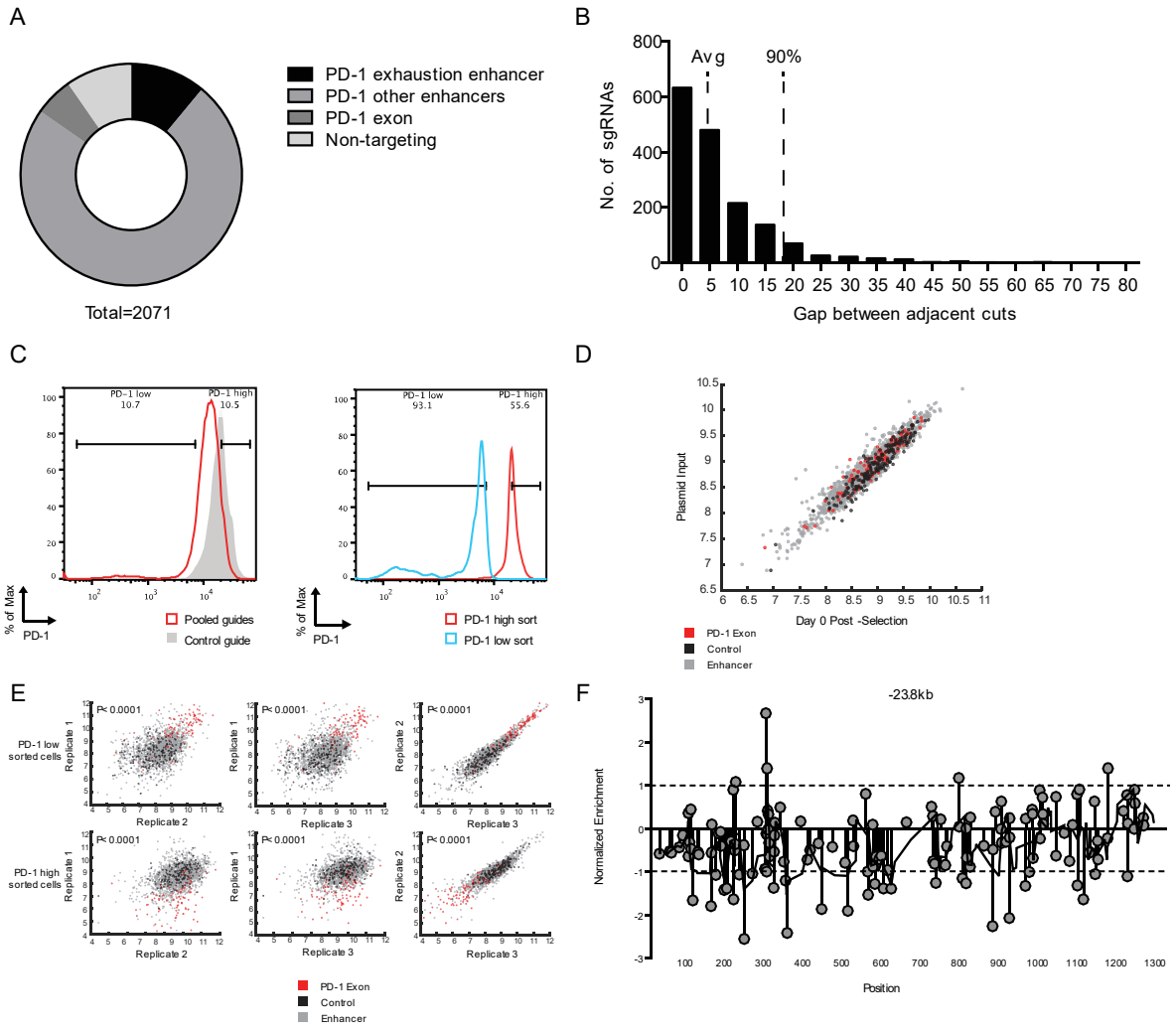
Supplementary Figure 3



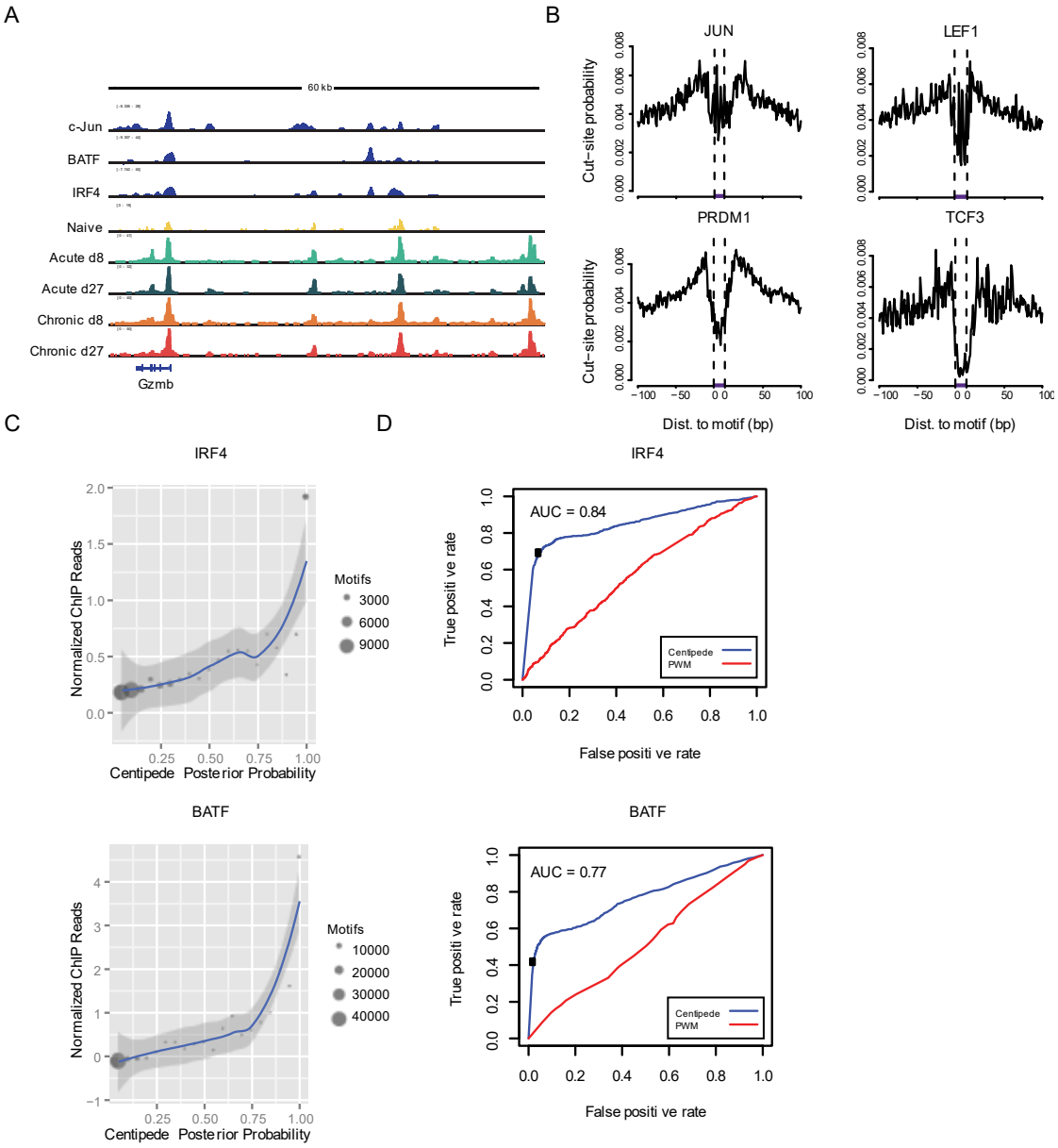
Supplementary Figure 4



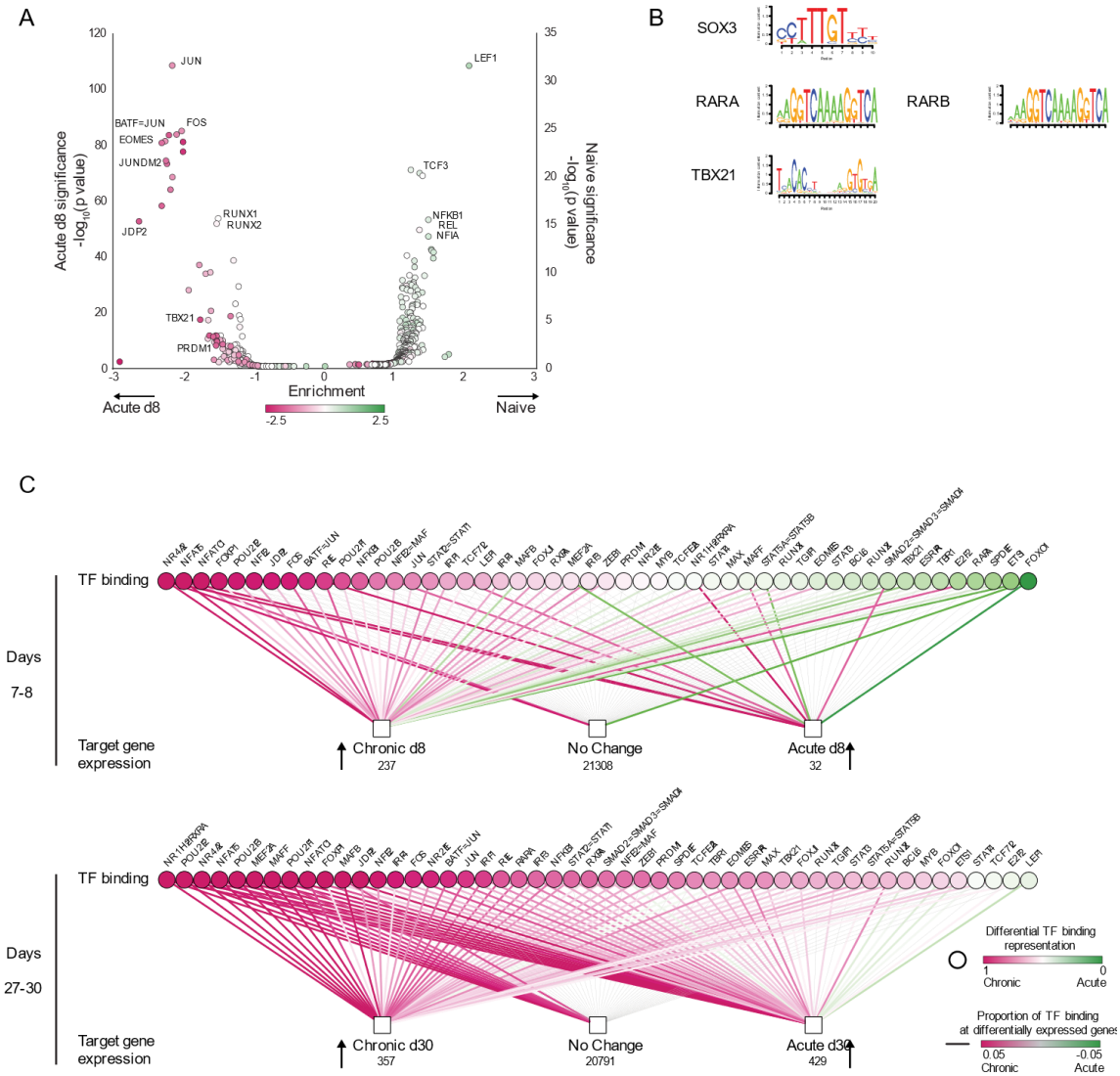
Supplementary Figure 5



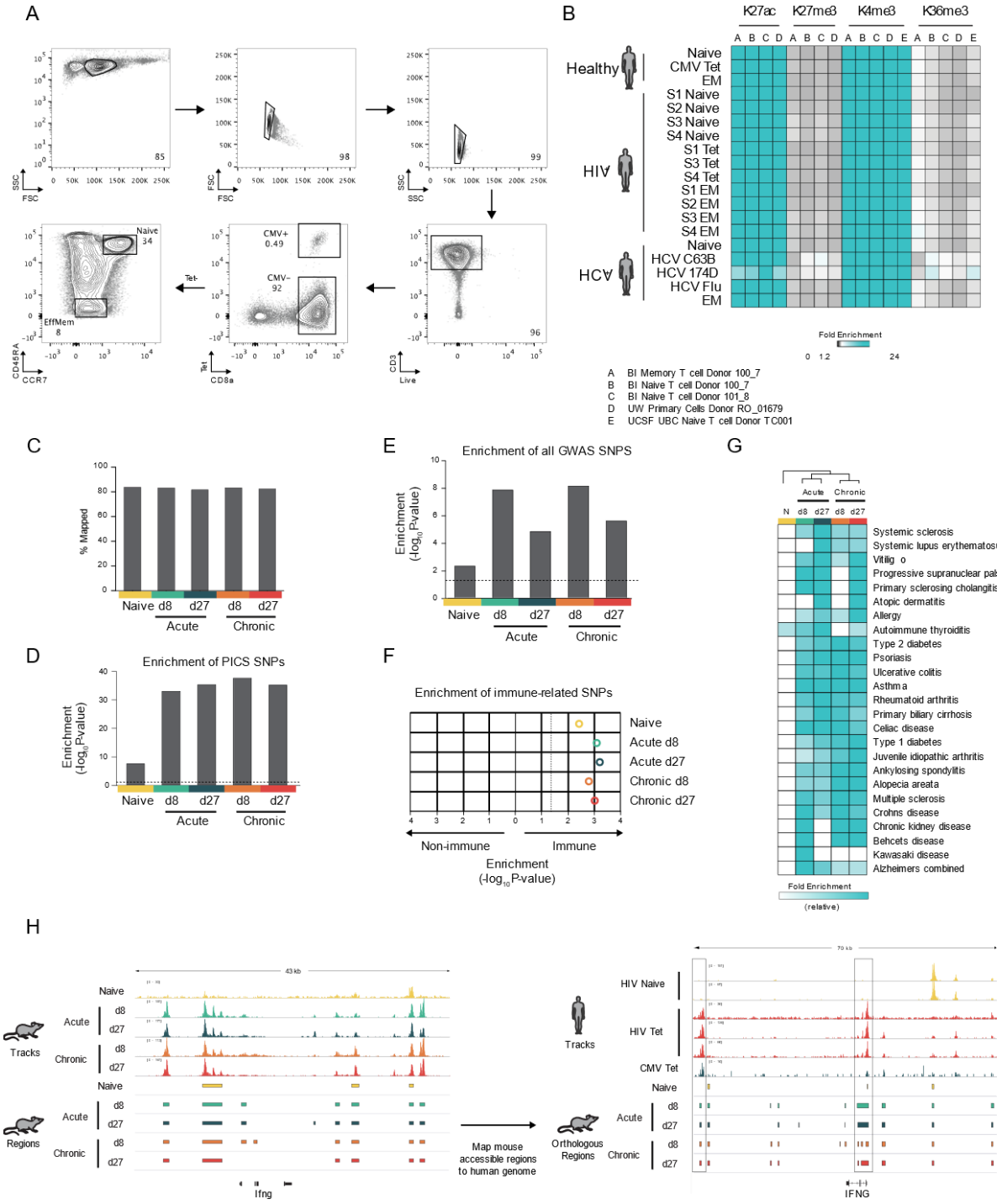
Supplementary Figure 6



Supplementary Figure 7



Supplementary Figure 8



Chapter 4 - Batf and Irf4 Regulate the Naive to Effector Trajectory

4.1 Introduction

In Chapter 3 of the dissertation, my co-authors and I presented evidence that chromatin accessibility changes between Naive, Effector, Memory, and Exhausted CD8+ T cells. Furthermore, we showed that the state-specific accessible regions are near genes with state-specific expression, and we validated a putative enhancer for PD-1 via CRISPR/Cas9 KO.

In Chapter 4, Hsiao-Wei Tsao, W. Nicholas Haining, Nir Yosef, and I investigate the regulation of chromatin accessibility and its impact on gene expression and phenotype by analyzing ATAC-Seq, ChIP-Seq, and RNA-Seq for wildtype and Batf knockout CD8+ cells (**Fig 1A**). Batf, "Basic leucine zipper transcription factor, ATF-like", regulates Th17 differentiation by typically co-binding with Irf4*,* and is also required for differentiation of other immune cell types (Glasmacher et al. 2012). (Kurachi et al. 2014) found that *Batf* is essential for differentiation of naive T cells to effector T cells, and that it simultaneously increases the expression of key effector TFs while reducing the expression of key effector products such as granzyme B.

We find that *Batf* and *Irf4* are essential for maintaining the regulatory landscape of effector T cells by altering the accessibility of the genome and the binding affinity of key transcription factors. We find evidence that *T-bet* and *Runx3* are highly enriched for binding in regions regulated by *Batf* and *Irf4*, and that these TFs are involved in long range chromatin interactions. Transduction and expression of these four genes in fibroblasts upregulated the expression several genes with important functions in T cells including cytokine receptors (*Il2rb* and *Il17ra*), a transcription factor (*Nfatc1),* a chemokine receptor (*Ccr7)*, proteins in cytokine signaling pathways (*Socs4* and *Jak1)* , cell surface receptors (such as *Cd300a)*, and others.

4.2 Knockout of BATF Downregulates Genes Critical for Effector Function

The naive to effector trajectory is initiated by TCR signaling. To determine which groups of genes are expressed during TCR stimulation, we obtained time course RNA-Seq data from *in vitro* artificially stimulated CD8+ T Cells. We applied hierarchical clustering to the gene expression data (**Fig 1B**), and found the key effector TFs *Batf, Irf4, and T-bet* (*Tbx21*) were mapped to Cluster 1 along with genes encoding key effector products such as *Ifng* and *Gzmb.* Genes in this cluster increase in expression during stimulation and restimulation. *Runx3*, another key TF, was assigned to Cluster 4, which contain genes which rise in the days after stimulation, but fall during restimulation.

To assess the impact of *Batf* knockout on gene expression, we obtained *in vivo* RNA-Seq data for WT and *Batf KO* P14 CD8+ T cells using a Cre-Lox system, and retrieved the cells

three and four days after LCMV infection (**Fig 1C**). Genes characteristic of Cytotoxic CD8+ T Cells such as *Ctla2* and *Ctla4* (cytotoxic T lymphocyte associated proteins 2a and 4) were downregulated in the *Batf KO* samples, and GO Enrichment analysis of genes that dropped more than twofold in expression between the WT and *Batf KO* samples revealed pathways relating to "T cell activation" and "regulation of leukocyte activation" (**Fig 1C**). Several genes were found in multiple pathways, such as *Lyn* (a protein tyrosine kinase), *Cd24a*, and *Ctla4* (**Fig1D**).

We obtained ATAC-Seq data from the same experimental setting as the *in vivo* Cre-Lox experiment and found several chromatin accessible regions (ChAR's) that were differentially accessible between the WT and *Batf KO* conditions (**Fig1E**). We mapped these ChAR's to putative target genes using GREAT (McLean et al. 2010) and then compared the average log2 fold change for Tn5 cuts in accessible regions to the log2 fold change in expression value for the target genes, focusing on genes with the most ChAR's mapped to them (**Fig1F**). Genes such as *Maf* (a bZIP transcription factor known to be expressed downstream of *Batf* and important for Tfh differentiation (Andris et al. 2017), and Th17 differentiation (Tanaka et al. 2014) [synonym: *c-Maf*]) and *Ctla4* (**Fig1G**) appear to have more differential chromatin regions nearby and higher gene expression in the WT vs *Batf KO*, while genes such as *Cd47* and *Tgfbr2* (part of the TGF-beta receptor) are upregulated in *Batf KO* condition. GREAT analysis of these differential peaks revealed an association between ChAR's lost in the *Batf KO* condition and the GO Biological pathway "negative regulation of T cell receptor signaling pathway**"**. Peaks gained in the *Batf KO* were associated with the pathway "positive regulation of T cell receptor signaling pathway". These results suggest that *Batf* regulates chromatin accessibility in regulatory regions that are key for processing the stimulation signal.

We also obtained ChIP-Seq data for *Batf* in WT *in vitro* cells that were restimulated six days after their initial stimulation. We identified 20,780 bound loci (ChIP-Seq) peaks for *Batf. Batf* binding overlapped with 3,666 of the 6,453 peaks that were differentially more open in the WT vs the *Batf KO* ATAC-Seq data. 3,486 of these 3,666 peaks were also bound by *Irf4*. GREAT analysis of the 20,780 Batf bound peaks (versus a background of 59,767 loci bound by Batf and other key regulatory factors) found enrichment for the pathways "T cell receptor signaling pathway" and "T cell receptor signaling pathway" (among others), and mapped Batf bound loci to genes such as the effector molecules *Ifng* and *Il2,* and transcription factors like *Jun* and various *Nfat*'s (Nuclear Factors of Activated T Cells). These results suggest that *Batf* may regulate the transcription of key effector genes through both changes in chromatin accessibility, and by binding near key genes without altering the chromatin landscape.

As mentioned earlier, the transcription factor *Irf4* is also known to co-bind with *Batf* and regulate important T cell functions. We produced ChIP-Seq data for *Irf4*, also in vitro CD8+ T Cells restimulated at Day 6, and identified 36,785 peaks, which covered 88.9% of the *Batf* bound ChIP-Seq peaks. We also carried out the same analysis of changes in chromatin accessibility and gene expression in WT vs *Irf4 KO* CD8+ T cells (**Fig 1J - 1M**) as

we did for WT vs *Batf* KO CD8+ T cells. Genes lost in the *Irf4* KO were enriched for GO pathways "regulation of leukocyte activation" and "regulation of leukocyte differentiation". ChARs lost in the *Irf4 KO* condition were mapped by GREAT to the pathway "negative regulation of T cell receptor signaling pathway". A comparison of the gene expression results (**Fig 1N**) and chromatin accessibility results (**Fig 1O**) for the WT vs *Batf* KO and WT vs *Irf4* KO finds similar effects for many genes and loci. 3,423 genes were significant in either comparison, and their estimated log2 fold changes had a Spearman correlation of .707. With respect to differentially accessibility, 15,007 peaks were differential in either comparison and their estimated log2 fold changes has a Spearman correlation of .539.

Several interesting immune-related genes display similar effects when either TF is knocked out. For example, the transcription factor *Maf* decreases in expression when either *Batf* or *Irf4* is disabled, and *Tlr7, a* Toll-like receptor increases in expression when either *Batf* or *Irf4* is removed. The similarity of *Batf* KO and *Irf4* KO on the chromatin landscape and gene expression, and their high degree of co-binding, implies that the two transcription factors work together to regulate effector function.

4.3 Critical CD8+ T Cell Transcription Factors Bind at Regions Regulated by BATF and IRF4

To understand what transcription factors regulate the changes we observed in gene expression, we obtained ChIP-Seq for ten key T cell transcription factors from *in vitro* P14 CD8+ T cells stimulated with anti-CD3, anti-CD28 and IL-2, and restimulated six days later with PMA+Ionomycin in WT and *Batf KO* conditions (**Fig 2A**). We assessed the co-binding of transcription factors (in the WT condition) by measuring the fold enrichment for binding (the percentage of "TF A" loci co-bound by "TF B", divided by the percentage of all Day6 TF Peaks bound by TF B, a symmetric measure) (**Fig 2B**). The TFs *Runx3* and *T-bet* exhibited high levels of co-binding with one another, and also with *Stat3* and *Stat5*. We also employed another measure of co-binding by assessing what percentage of a TF's regions were co-bound by other TFs (**Fig 2C**). TFs such as *JunD*, which are bound to many loci, also co-bind many of the sites bound by the other key TFs. TFs which bind to far fewer loci, like *T-bet* and *Runx3*, are covered by many of the other TFs. We investigated co-binding relationships within the ChAR's that are differentially closed in the BATF KO condition (**Fig 2D**), and find that *T-bet* and *Runx3* are highly enriched. Many of the other TFs also show enrichment for binding in these loci, implying they may have some regulatory function.

We then identified regions which were differentially bound by TFs between the WT and *Batf KO* conditions (**Fig 2E**). The transcription factors *JunD*, *Stat5*, and *T-bet* mainly lost binding, and the TFs *Eomes*, *Fosl2*, and *Stat3* mainly gained binding. To investigate whether changes in binding were due to changes in chromatin accessibility or a co-binding relationship with *Batf*, we compared the fold change in ATAC-Seq to the fold change for ChIP-Seq at bound loci. Transcription factors such as *JunD*, *Stat5*, *Irf4*, and *T-*

bet generally lost both accessibility and binding in the regions where Batf is bound, whereas TFs such as Ets1 and Stat3 do not have a strong relationship between accessibility and co-binding in the regions where they are bound with Batf. This suggests that *Batf* may affect the binding of some transcription factors through accessibility, while impacting others by co-binding and altering their binding affinity.

4.4 HiChIP Data Reveals Long Range Contacts Between Key Regulatory Regions

We sought to understand whether *Batf* is associated with long range contacts between CD8+ T cell regulatory regions, and carried out the HiChIP protocol on CD8+ restimulated Day 6 T cells (**Fig 3A, Fig3B**) for *Batf* and *Ctcf*. Fold enrichment analysis of the endpoints/anchors of HiChIP revealed higher enrichment for binding by our key transcription factors and differentially open/closed ChAR's compared to *Ctcf* (**Fig 3C**). Furthermore, we see that fold enrichment increases with the minimum loop length, and we see that *T-bet* and *Runx3* score particularly high on this measure.

We assessed whether particular TFs and classes of differential ChAR's are more likely to interact at long distances by examining the percentage of peaks for these individual features that are connected to other features at long distance loops (**Fig 3D**). Relative to the other transcription factors, *Runx3* and *T-bet* come into contact with other TFs at higher rates. This is commensurate with research that suggests *T-bet* may be able to link distal regions of the genome (Liu et al. 2016) .

4.5 Forced Expression of *Batf*, *Irf4*, *T-bet*, and *Runx3* Shifts Transcriptome of Fibroblasts to be More Similar to T Cells

Our previous analyses showed that knockout of *Batf* and *Irf4* causes substantial changes to gene expression and chromatin accessibility. Analysis of ChIP-Seq in these regulatory regions and in HiChIP data displayed high enrichment for *Runx3* and *T-bet*. We hypothesized that these four transcription factors may play a crucial role in regulating CD8+ T cell function, and transfected and expressed them in fibroblast cells and used RNA-Seq and ATAC-Seq to investigate their impact on gene expression and chromatin accessibility. We tested all sixteen possible combinations of the four genes.

RNA-Seq analysis of the differential genes (**Fig 4A**) finds increased expression of *Il2rb*, *Ccr7*, *Il17ra*, *Nfatc1*, *Socs4*, *Jak1*, and *Cd300a* (among other genes). Differential ATAC-Seq peaks reveals opening of peaks near *Il2rb, Ccr7, Il17ra, Nfatc1, Cd300a, and Jak1*. Several of these genes fulfill important T cell functions, and include cytokine receptors (*Il2rb* and *Il17ra*), a transcription factor (*Nfatc1),* a chemokine receptor (*Ccr7)*, proteins in cytokine signaling pathways (*Socs4* and *Jak1)*, and cell surface receptors (such as *Cd300a).*

We also conducted HiChIP on fibroblast samples, adding *Batf* alone to one sample, and *Batf*, *Irf4*, *T-bet*, and *Runx3* to a second sample. While both samples were less enriched

for T cell regulatory regions than our HiChIP T cell samples, the fibroblasts with all four transcription factors are slightly more enriched than those with just *Batf* (**Fig 4D**).

To assess whether the overall gene expression of the fibroblasts became more similar to T cells as our four key transcription factors were added, we created two "T Cell signatures" by comparing the expression of Naive T cells and Effector T cells to unmodified fibroblasts (**Fig 5A and 5B**). (The four TFs were excluded from the signatures so that their particular expression would not bias the results.) On both measures, the fibroblasts with transcription factors added generally scored higher on the T Cell signatures (**Fig 5C**), with fibroblasts expressing both *Batf* and *Irf4* typically exhibiting the strongest T Cell scores. We computed the Shapley value for each TF (**Fig 5D**), a measure of the individual TF's contribution to the increase/decrease in gene signature score, and found that *Batf* and *Irf4* were the highest contributors.

Interestingly, *Runx3* has a negative Shapley Value, implying that the transcription factor is shifting the cells away from our three T Cell fates. *Runx3* is known to upregulate the expression of cytotoxic molecules, and has also been shown to repress expression of genes that can direct cells towards other T Cell fates, like *Tcf7 and Cxcr5* which lead towards the Follicular Helper T Cell fate (Shan et al. 2017). *Tcf7* and *Cxcr5* receive a score of +1 in both T cell signatures, so perhaps some of the negative Shapley value is due to *Runx3* repressing genes like these.

We conducted a similar analysis using the chromatin accessibility data in place of the gene expression data (**Figures 5E - 5H**), and found that *Irf4* had the greatest contribution to shifting the chromatin landscape to a more T cell-like state. *Batf* and *T-bet* had the next greatest contribution, followed by *Runx3* which had a small positive effect, as opposed to the small negative effect displayed in the RNA-Seq results. As in the RNA-Seq analysis, the samples expressing both *Batf* and *Irf4* tend to be the most T cell-like (with the exception of the *Batf+Irf4+Runx+Tbet-* datapoint), suggesting that *Batf* and *Irf4* are both necessary to shift chromatin to a more T cell like state.

4.6 Conclusions

Our results show that *Batf* and *Irf4* are essential for maintaining the regulatory landscape for effector function by remodeling chromatin and allowing key transcription factors to bind. We find that *T-bet* and *Runx3* are particularly enriched in these regions. Both TFs are enriched for long-range contacts, commensurate with evidence in the literature that *T-bet* may be a chromatin linker. Furthermore, expression of these four transcription factors in fibroblasts upregulates the expression of several characteristic T cell genes.

4.7 Methods

Note: Dr. Hsiao-Wei Tsao, my colleague on the project who carried out the experiments, wrote the sections, ***In Vivo*** **T Cell Preparation** and ***In Vitro*** **T Cell Preparation,** and edited the sections **Fibroblast Transfection and Preparation** and **Preparation of HiChIP Samples.**

In Vivo T Cell Preparation

BATF$^{f/f}$ x CreERT2 x P14, IRF4$^{f/f}$ x CreERT2 x P14, CreERT2 x P14, and P14 control donor mice (with CD45.2+ congenic marker) were treated with 2 mg 4-OHT for 5 consecutive days. Spleens were harvested 5 days after final 4-OHT. P14 cells were enriched by CD8 negative selection and analyzed prior to transfer. The enriched CD8+ T cells were then transferred to the CD45.1+ B6 recipient mice One day after the transfer, the recipient mice were infected with the lymphocytic choriomeningitis virus (LCMV) strain Armstrong. Three and four days post infection, the spleens were harvested, LCMV-specific CD8+ T cells were sorted, and processed with the ATAC-Seq and RNA-Seq protocols.

In Vitro T Cell Preparation

CD8+ T cells were isolated from P14 B6 mice with CD8a+ T cell isolation kit (Milteny Biotec, 130-104-075) and stimulated at Day 1 with anti-Cd3e (BD, clone 145-2C11, 2 ug/ml precoated) and 2 ug/ml anti-Cd28 (BD, clone 37.51) with complete RMPI medium supplemented with 100 U/ml recombinant human IL-2. Cells were expanded and cultured in vitro with complete RPMI and IL-2 for 6 days. Cells marked as “restimulated were stimulated again at Day 6 with 50 ng/ml PMA and 1 uM ionomycin for three hours. Cells were processed in the RNA-Seq protocol (NEBNext Ultra RNA Library Prep Kit, E7530 or NEBNext Ultra RNA Library Prep Kit, E7770) at Days 0 (Naive), 1, 2, 3, 4, 5, 6, and again at Day 6 after restimulation. Cells for ChIP-Seq were processed at Day6, with restimulation and fixed with 1% formaldehyde at 37 degree for 10 mins. Control samples without an antibody were also submitted in the the ChIP-Seq sequencing batches.

Fibroblast Transduction and Preparation

NIH 3T3 murine fibroblast cells were transduced with Doxocyline inducible lentivirus expressing *Batf, Irf4, Runx3,* and *T-bet* with hygromycin B, Blasticidin, Puromycin and GFP as selection markers, respectively. The transduced fibroblast cells were selected with proper antibiotics for more than a week and treated with 4 ug/ml doxycycline for additional 72 hours before harvested. All sixteen possible combinations of these TFs (including samples without any of the four TFs) were prepared for ATAC-Seq, and RNA-Seq, and the Batf and Batf+Irf4+Tbet+Runx3 samples were prepared for HiChIP (fixed with 1% formaldehyde). For each combination of TFs prepared for ATAC-Seq and RNA-Seq, two replicates received Doxocycline treatment, and two replicates did not.

Preparation of HiChIP Samples

Ten million fixed cells were used for each HiChIP sample preparation. Two HiChIP samples from T Cells were prepared using *Ctcf* (Millipore, 07729) and *Batf* (Brookwood biomedical, PAB4003) as the targets according to the protocol described in (Mumbach et al. 2016). A fibroblast sample with *Batf* expressed, and a fibroblast sample with *Batf*, *Irf4*, *T-bet*, and *Runx3* expressed were also processed according to the same protocol.

Alignment of RNA-Seq

All RNA-Seq reads were trimmed using Trimmomatic (Bolger, Lohse, and Usadel 2014) to remove primer and low-quality bases. Reads were then passed to FastQC [http://www.bioinformatics.babraham.ac.uk/projects/fastqc/] to check the quality of the trimmed reads. The single-end reads were then aligned to the Mus_musculus.GRCm38.82 transcriptome from Ensembl using RSEM (Li and Dewey 2011) with the parameters "--num-threads 4 --bowtie2 --sampling-for-bam --output-genome-bam --sort-bam-by-coordinate --sort-bam-memory-per-thread 1G --estimate-rspd --fragment-length-mean 200 --fragment-length-sd 80'.

Analysis of Differential Gene Expression

Differential expression was assessed using DESeq2 (Love, Huber, and Anders 2014) . Genes were considered differentially expressed if they had an adjusted pvalue <0.05. In models without an interaction term, "betaPrior" was set equal to TRUE, so that an expanded model matrix could be retrieved. In models with an interaction term, "betaPrior" was set to FALSE, to allow for use of the interaction term.

Alignment of ATAC-Seq and Peak Calling

All ATAC-Seq reads were trimmed using Trimmomatic (Bolger, Lohse, and Usadel 2014) to remove primer and low-quality bases. Reads were then passed to FastQC [http://www.bioinformatics.babraham.ac.uk/projects/fastqc/] to check the quality of the trimmed reads. The paired-end reads were then aligned to the mm10 reference genome using bowtie2 (Langmead and Salzberg 2012), allowing maximum insert sizes of 2000 bp, with the "--no-mixed" and "--no-discordant" parameters added. Reads with a mapping quality (MAPQ) below 30 were removed. Duplicates were removed with PicardTools , and the reads mapping to the blacklist regions and mitochondrial DNA were also removed. Reads mapping to the positive strand were moved +4 bp, and reads mapping to the negative strand were moved -5bp following the procedure outlined in (Buenrostro et al. 2013) to account for the binding of the Tn5 transposase.

Peaks were called using macs2 on the aligned fragments (Zhang et al. 2008) with a qvalue cutoff of 0.05 and overlapping peaks among replicates were merged. A region was considered a valid peak if it had a qvalue below 0.05 in at least two replicates, and a qvalue below 0.001 in at least one of the replicates.

Tests of Differential Accessibility in ChAR's

Differential accessibility was assessed using DESeq2 (Love, Huber, and Anders 2014) with a matrix of "peaks by samples" replacing the "genes by samples" matrix. Counts of Tn5 cuts were used instead of gene expression values. Peaks were considered differentially accessible if they had an adjusted pvalue <0.05. In models without an interaction term, "betaPrior" was set equal to TRUE, so that an expanded model matrix could be retrieved. In models with an interaction term, "betaPrior" was set to FALSE, to allow for use of the interaction term.

Alignment of ChIP-Seq and Peak Calling

All ChIP-Seq reads were trimmed using Trimmomatic (Bolger, Lohse, and Usadel 2014) to remove primer and low-quality bases. Reads were then passed to FastQC [http://www.bioinformatics.babraham.ac.uk/projects/fastqc/] to check the quality of the trimmed reads. These single-end reads were then aligned to the mm10 reference genome using bowtie2 (Langmead and Salzberg 2012), allowing maximum insert sizes of 2000 bp, with the "--no-mixed" and "--no-discordant" parameters added. Reads with a mapping quality (MAPQ) below 30 were removed. Duplicates were removed with PicardTools , and the reads mapping to the blacklist regions and mitochondrial DNA were also removed.

Each combination of transcription factor and condition had two replicates, with the exception of Runx3-WT, which had three replicates. ChIP-Seq peaks were called in each replicate, versus a control sample, using macs2 (Zhang et al. 2008) with a qvalue cutoff of 0.05 and overlapping peaks among replicates were merged. A region was considered a valid peak for a transcription factor if it had a qvalue below 0.05 in at least two replicates, and a qvalue below 0.001 in at least one of the replicates.

Tests of Differential Binding

Differential binding was assessed using the csaw package (Lun and Smyth 2016) in regions identified as being bound in either the WT or Batf Ko for the given transcription factor. Only log2 fold changes greater than one with a qvalue below 0.05 were considered differentially bound.

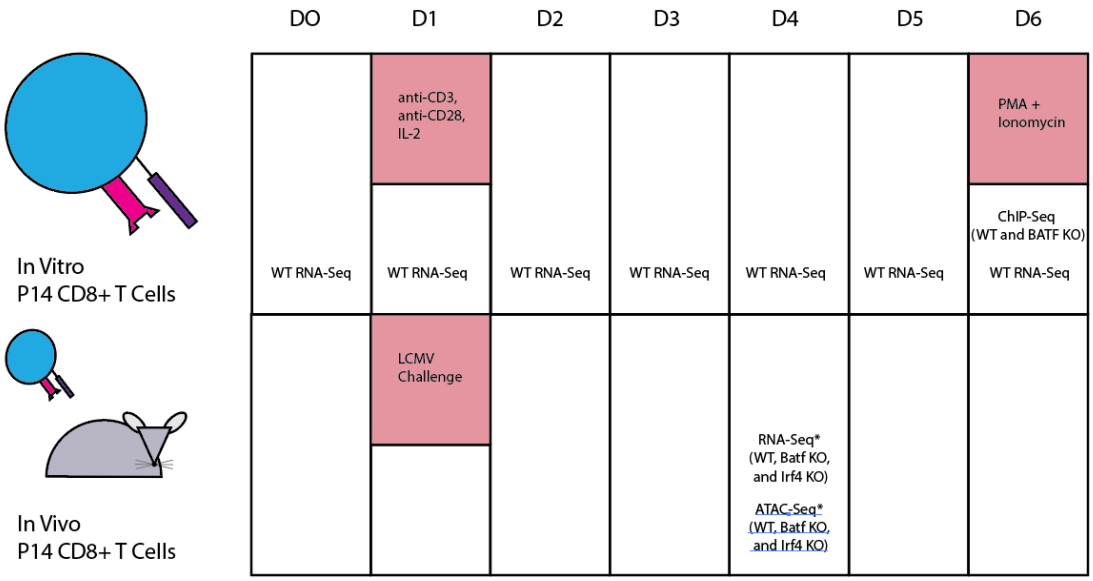
Alignment of HiChIP Data and Peak Calling

HiChIP data were aligned using the HiCPro pipeline (Servant et al. 2015). Normalized counts were plotted using the HiCPLotter python package (Kadir Akdemir - https://github.com/kcakdemir/HiCPlotter).

The Hichipper package (Lareau and Aryee 2018) was used to identify loops in the HiChIP data. The package relies on the Mango package (Phanstiel et al. 2015) to identify statistical significance of loops.

4.7 Figures

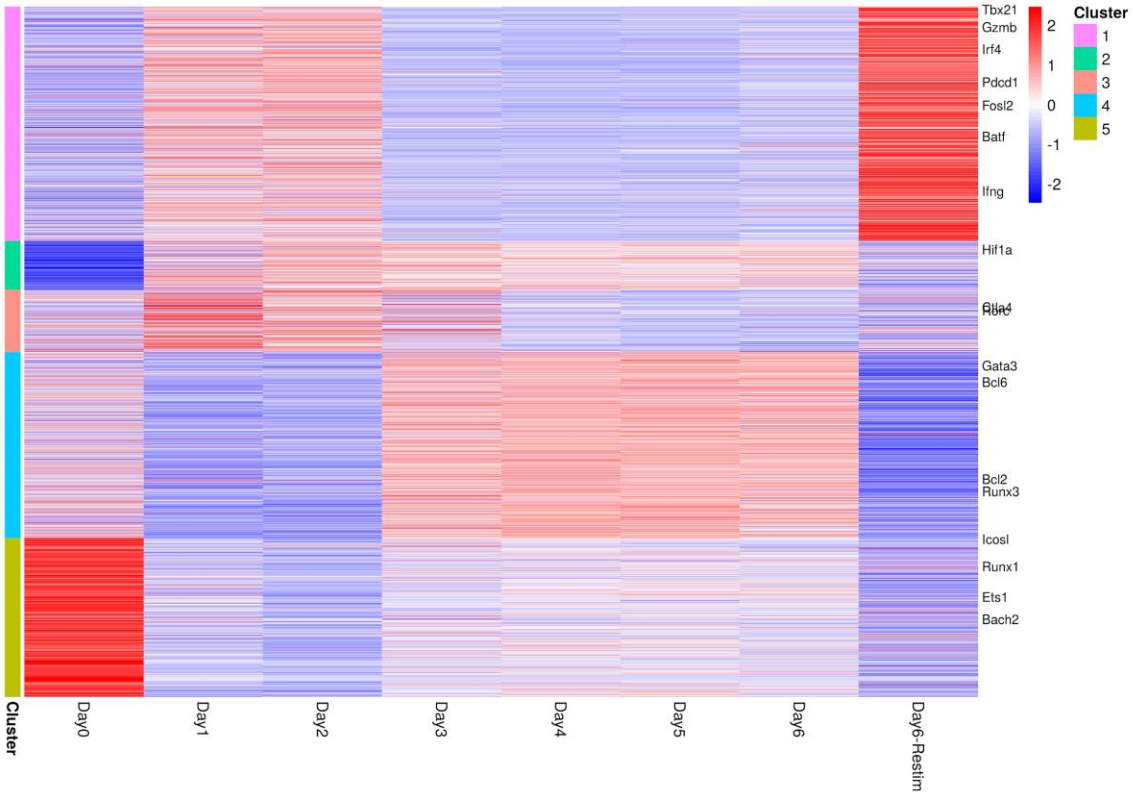
Figure 1A - Gene Expression, Chromatin Accessibility, and Transcription Factor Binding Data for CD8+ T Cells



Legend: The schematic above depicts the timepoint, experimental setting, and cell type for our initial data.

* The cells for this sample were harvested at both Day3 and Day4. A Cre-Lox system was used to KO the gene of interest.

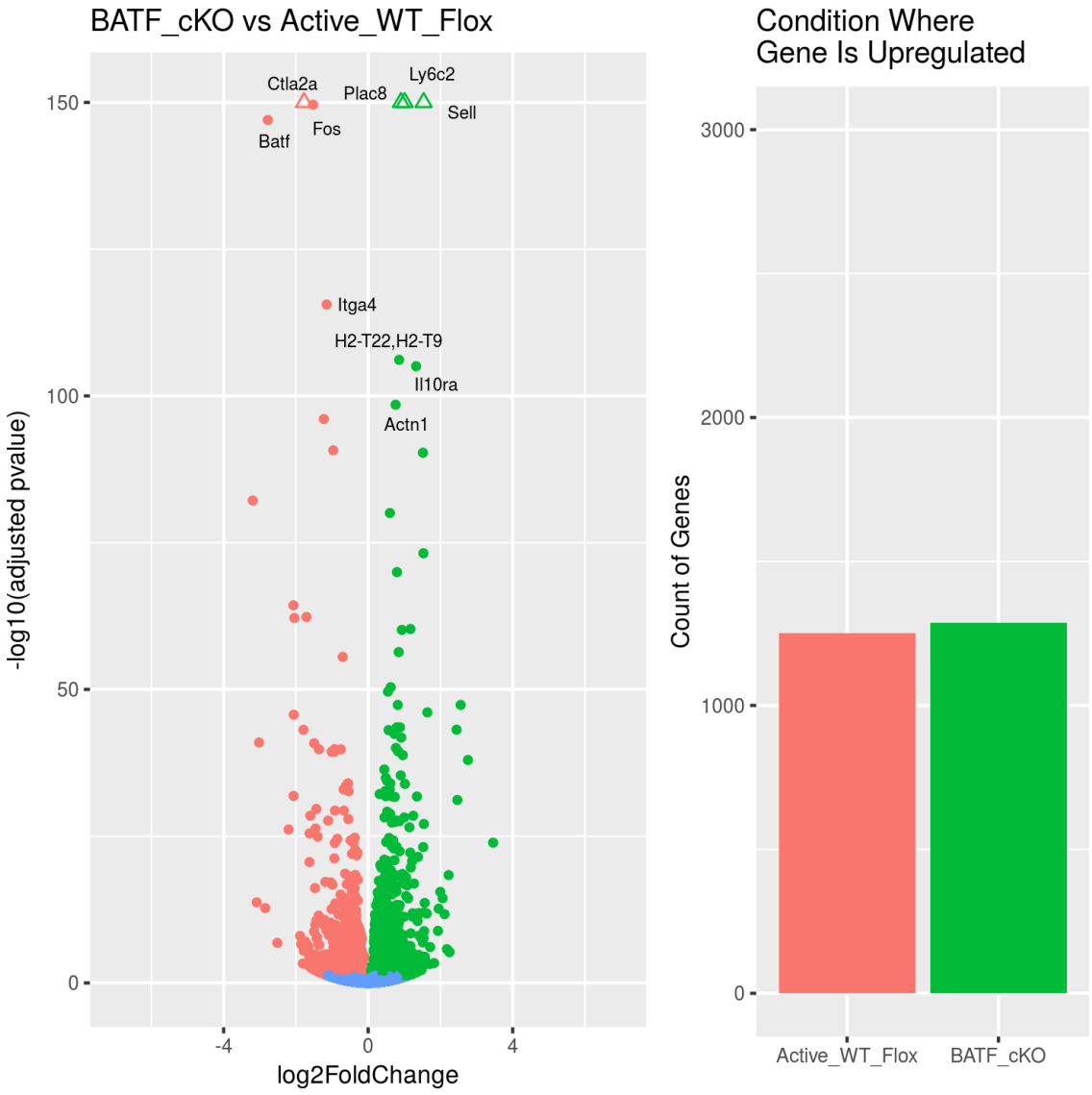
Figure 1B - Genes Display Distinct Patterns of Expression in Response to Stimulation



Legend: Each row is a gene, each column is a timepoint, and the cell represents the average expression of a gene across three *in vitro* replicates for that timepoint. "Day6" cells were not stimulated again; "Day6-Restim" samples were restimulated with PMA and Ionomycin.

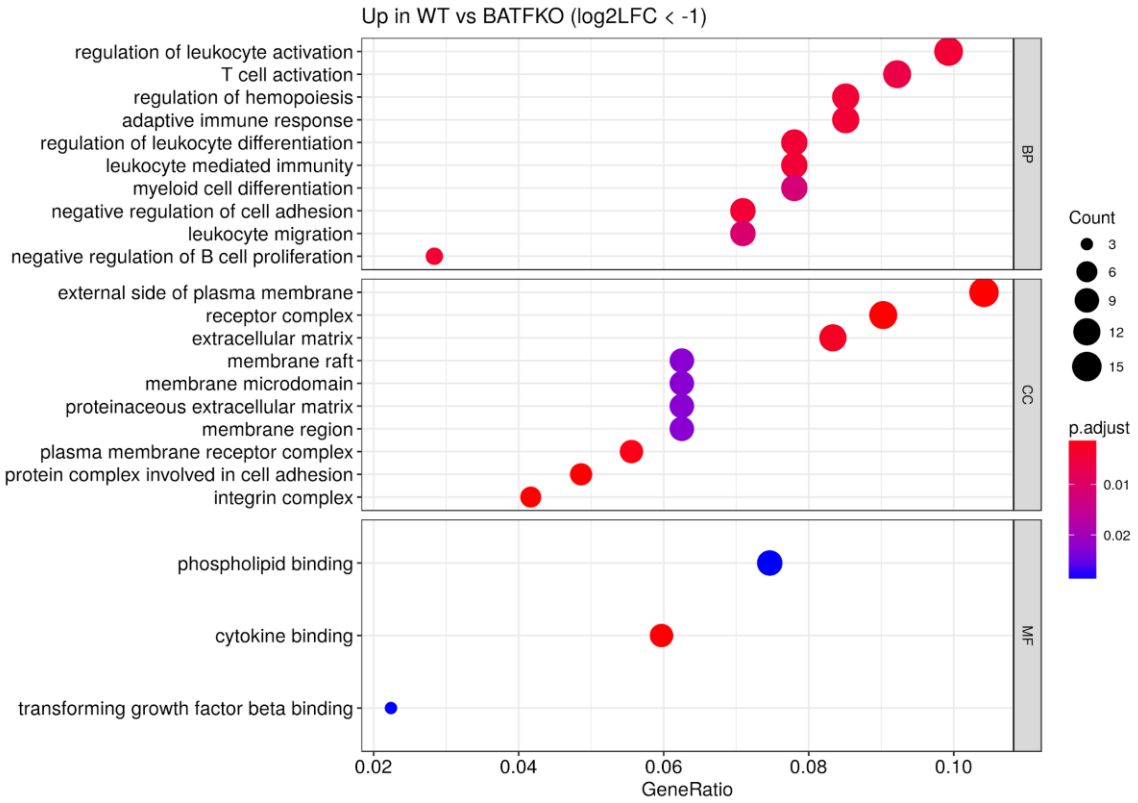
DESeq2 was used to identify genes that change with respect to the timepoint/stimulation categories by applying a likelihood ratio test, and only genes with an adjusted pvalue <0.05 were used. Normalized gene expression values were obtained from DESeq2 for each replicate. The mean values of these replicates were row normalized, and clustered by Pearson correlation using the pheatmap R package.

1C - Knockout of *Batf* Changes Transcriptome of CD8+ T cells



Legend: The volcano plot on the left indicates genes that were differentially expressed in the "WT Flox" (red) and "Batf cKO" (green) conditions. Blue genes were not differentially expressed. The barplot indicates the number of differentially expressed genes found in each condition. (Genes with an adjusted pvalue <0.05 are considered differentially expressed.)

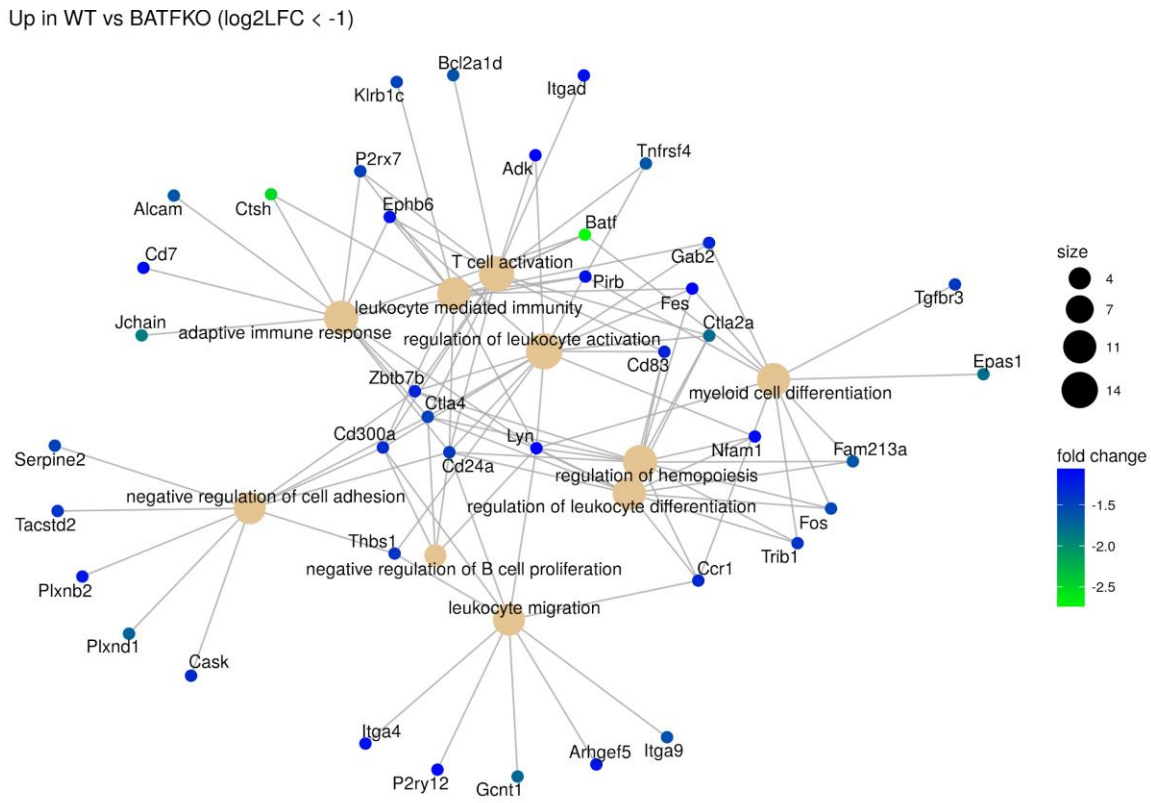
1D - Genes Strongly Upregulated in the WT vs the *Batf* KO Condition are Enriched for Pathways Regulating T Cell Activation, Differentiation and Other Key Functions



Legend: Top-scoring significant GO pathways by hypergeometric test (enrichGO in clusterProfiler package) for genes that were differentially upregulated at least twofold in the WT vs *Batf KO* test (n=156) . GO Pathways are categorized into Biological Process (BP), Cellular Component (CC), and Molecular Function (MF). “Gene ratio” represents the percentage of genes in the “differentially up” set that were also in the GO pathway.

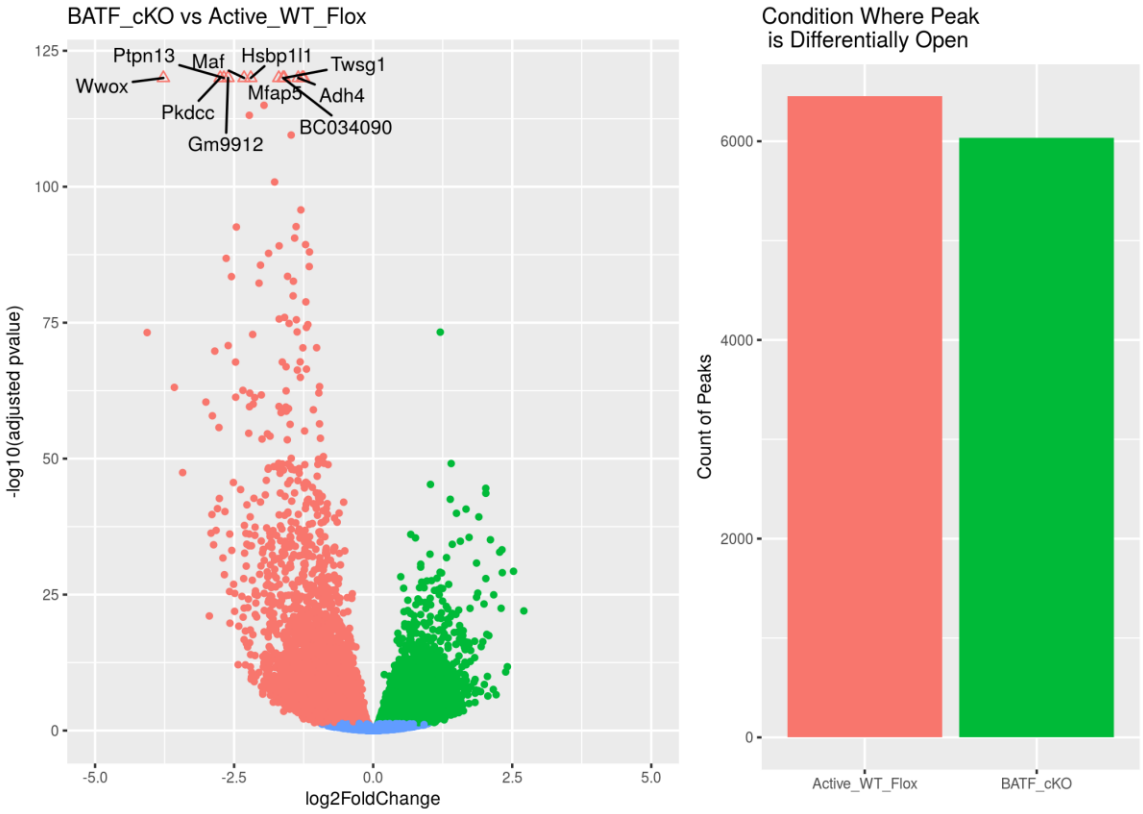
The GO enrichment test and plot were carried out with the “clusterProfiler” R package. (Yu et al 2012)

Fig 1E - Several Genes Lost in the *Batf KO* Condition are Shared Across Key Immune Pathways



Legend: Each beige point represents one of the top ten networks (ranked by pvalue) that were strongly upregulated (fold change >=2) in the WT vs *Batf* KO condition. The GO enrichment test and plot were carried out with the "clusterProfiler" R package. (Yu et al 2012) Note: The "fold change" displayed here is the log2 Fold Change estimate from DESeq2.

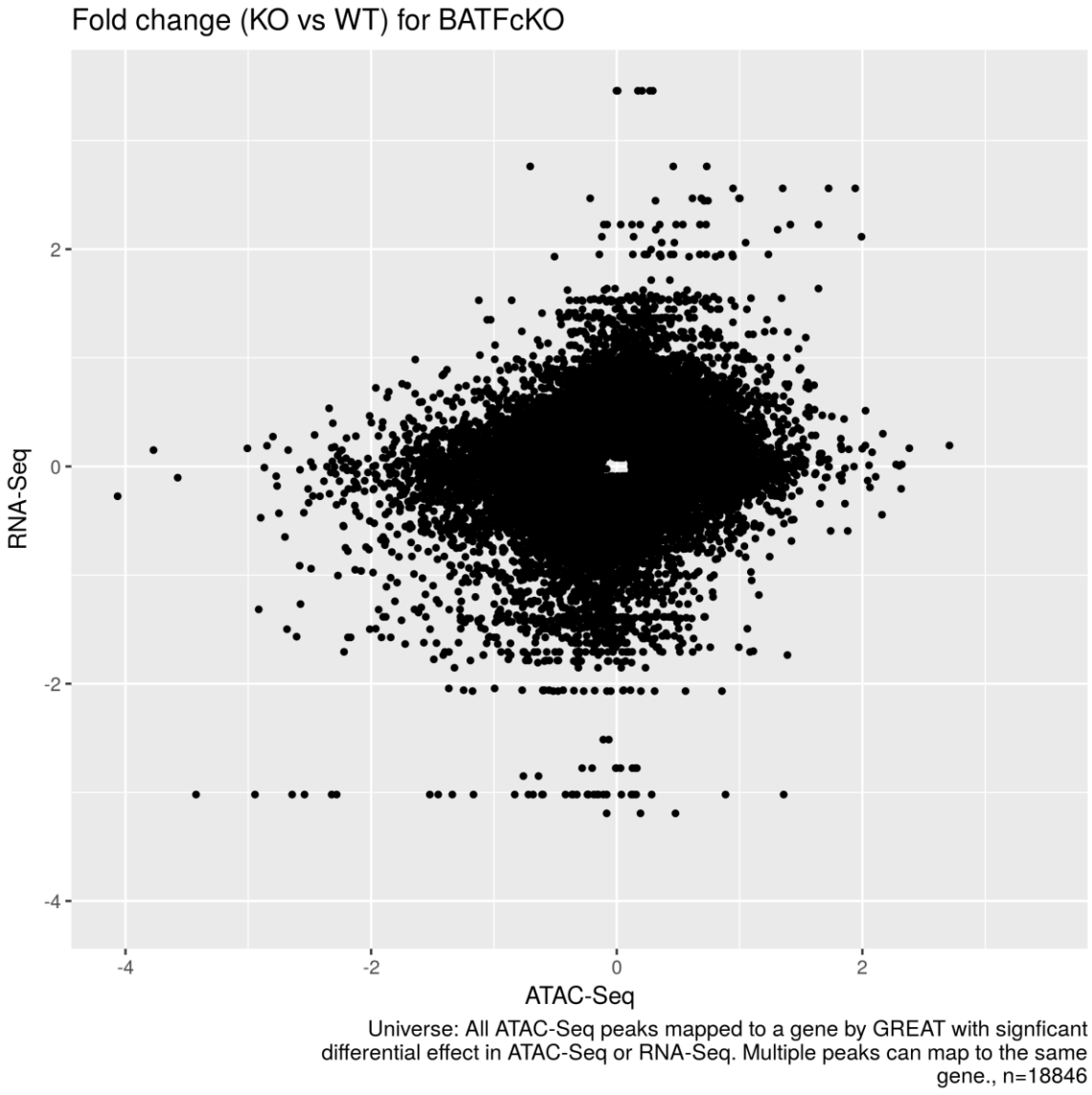
Fig 1F - Knockout of *Batf* Results in Extensive Remodeling of Chromatin Landscape



Legend: The volcano plot on the left indicates regions on the genome that were differentially accessible in the "WT Flox" (red) or "Batf cKO" (green) conditions. Blue loci were not differentially accessible. The barplot indicates the number of differentially accessible regions found in each condition. (Regions with an adjusted pvalue <0.05 in the test are considered differentially accessible.)

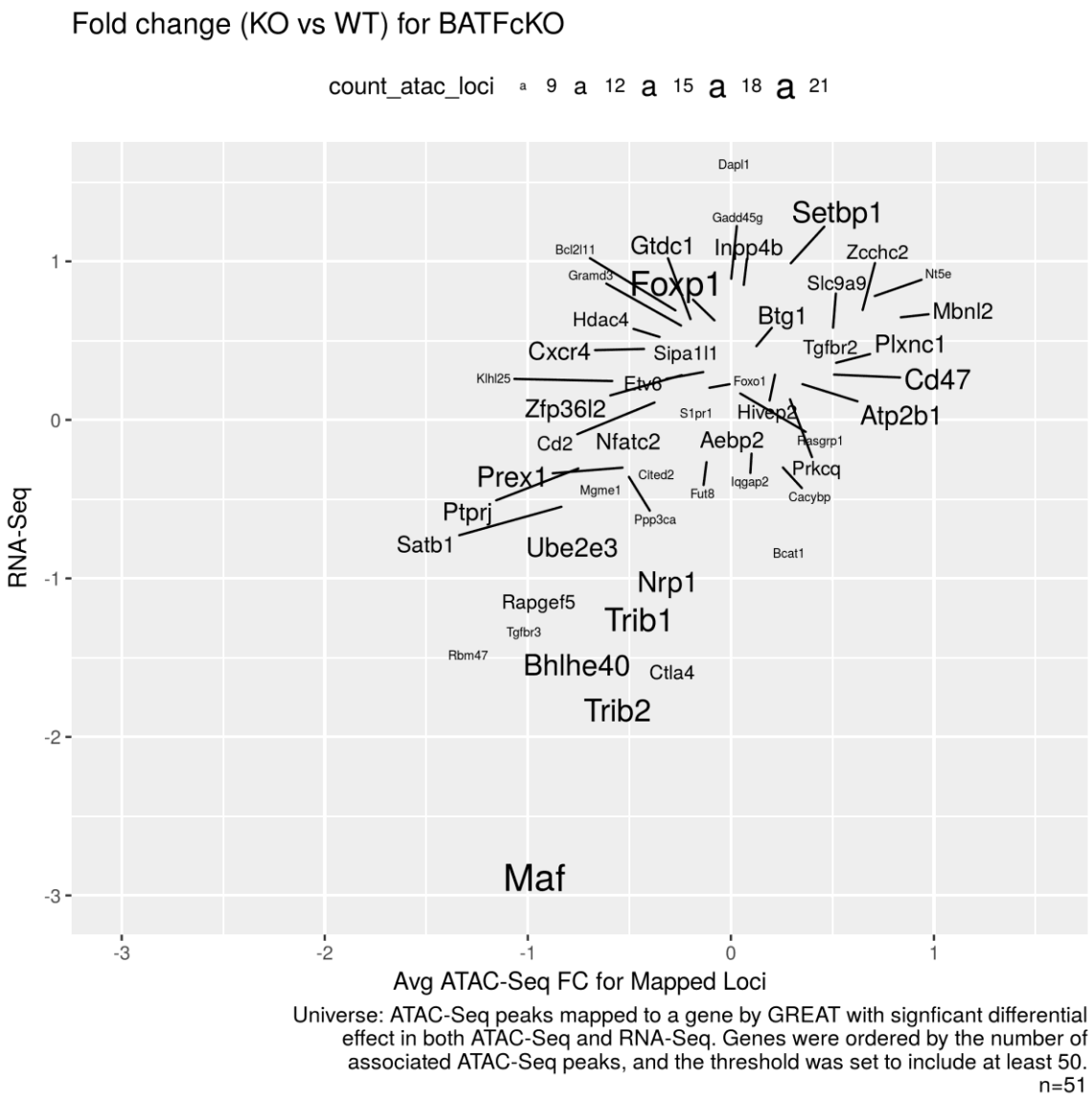
GREAT (McLean et al 2000) was used to map loci to genes they may regulate, and the closest gene for the most differential loci are displayed.

Fig 1G - Changes in Chromatin Accessibility and Expression of Nearby Genes Vary in *Batf* KO



Legend: The X-axis represents the log2(Batf_KO/WT) ratio of ATAC-Seq Tn5 cuts, and the y-axis represents the log2(BatfKO/WT) ratio of gene expression for the corresponding gene. Multiple ChAR's can map to one particular gene. The log2 ratios were estimated by DESeq2.

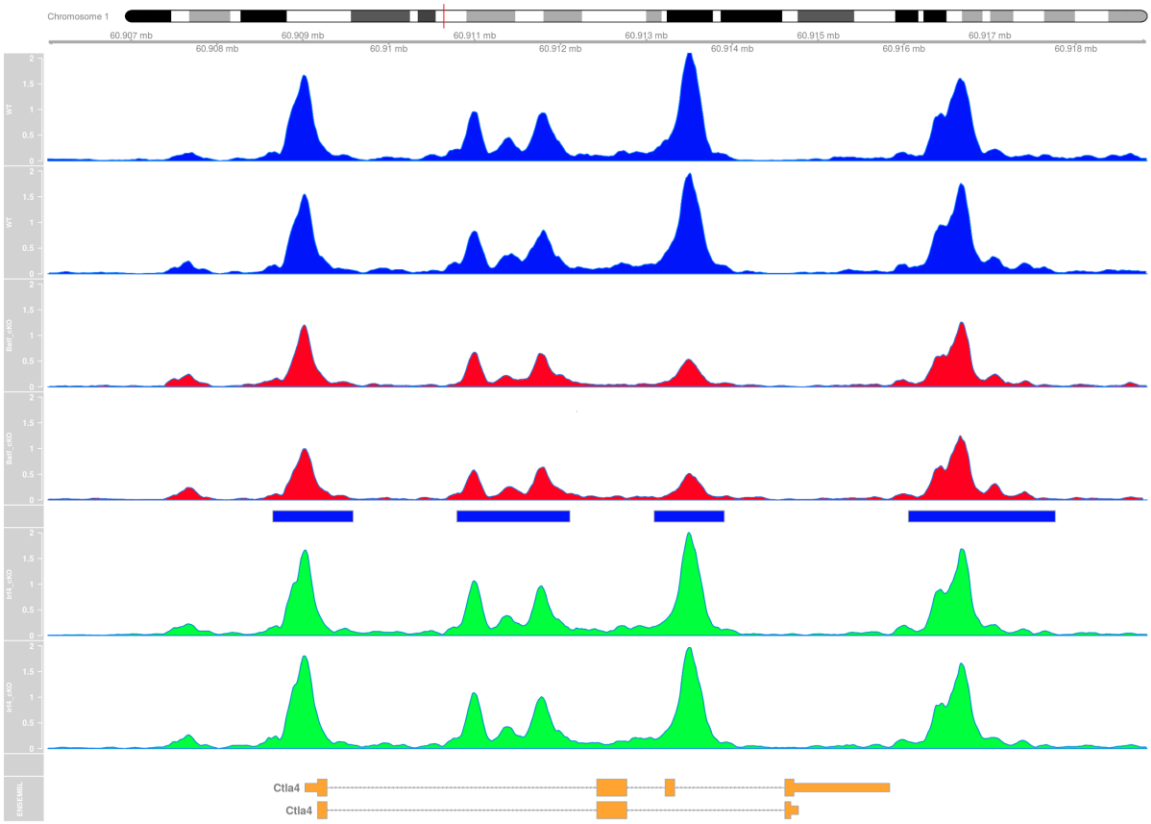
Fig 1H - Several Key CD8+ T Cell Genes Show Remodeling of Nearby Chromatin in Addition to Changes in Gene Expression in *Batf* KO.



Legend: Plot of genes that had both differential expression and multiple differential ChAR's (Chromatin Accessible Regions) mapped to them by GREAT. The size of gene name represents the number of differential ChAR's mapped to it.

Genes with the most differential ChAR's mapped to them were plotted, and the threshold was set to include at least 50.

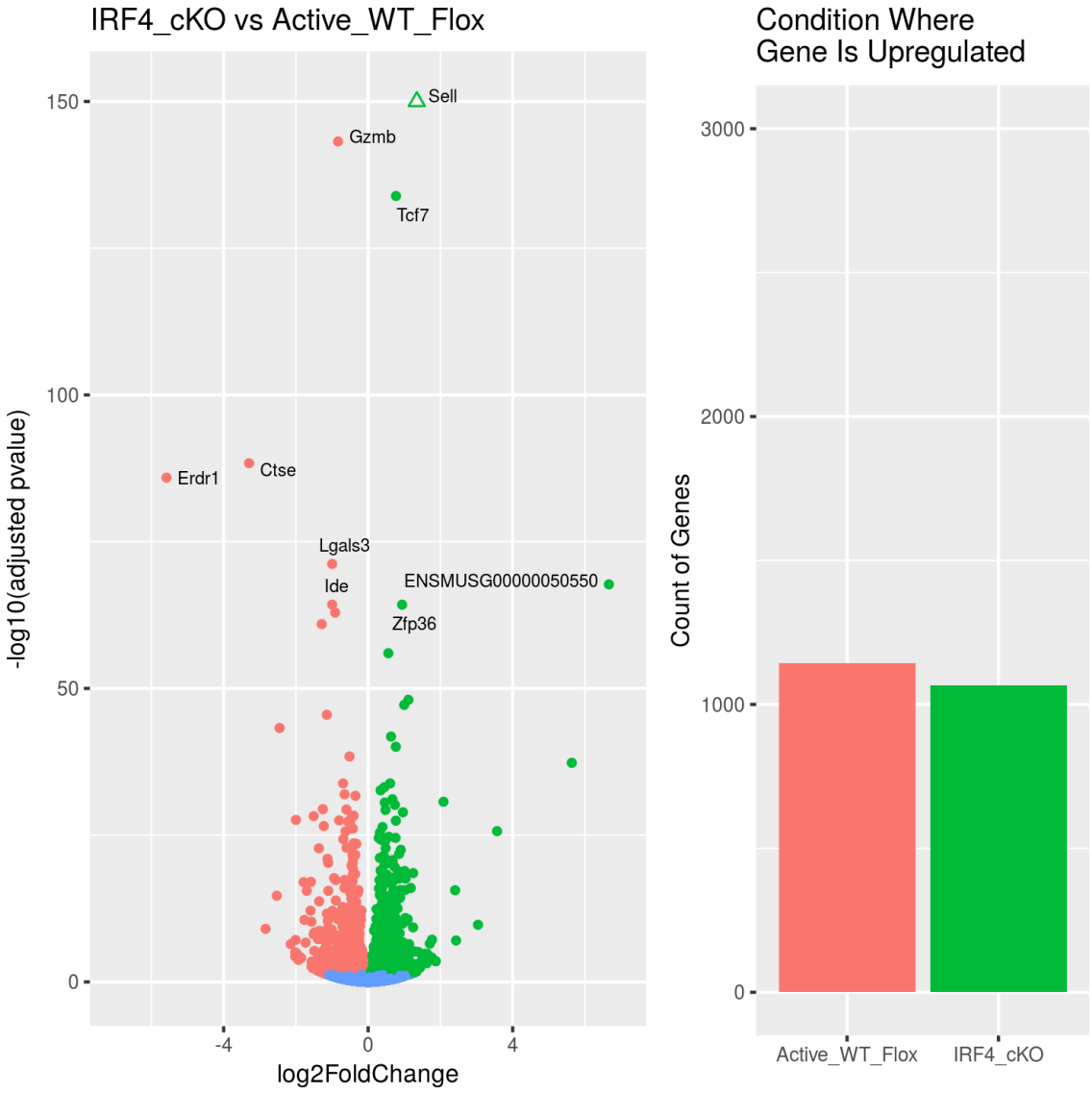
Fig 1I - Chromatin Remodeled Near *Ctla4* by Batf KO



Legend: Plot of ATAC-Seq fragments from replicates for the WT (blue), Batf KO (red), and Irf4 KO (green) near *Ctla4*. Blue rectangles indicate regions where DESeq2 identified a differential increase in accessibility for the WT condition versus the Batf KO.

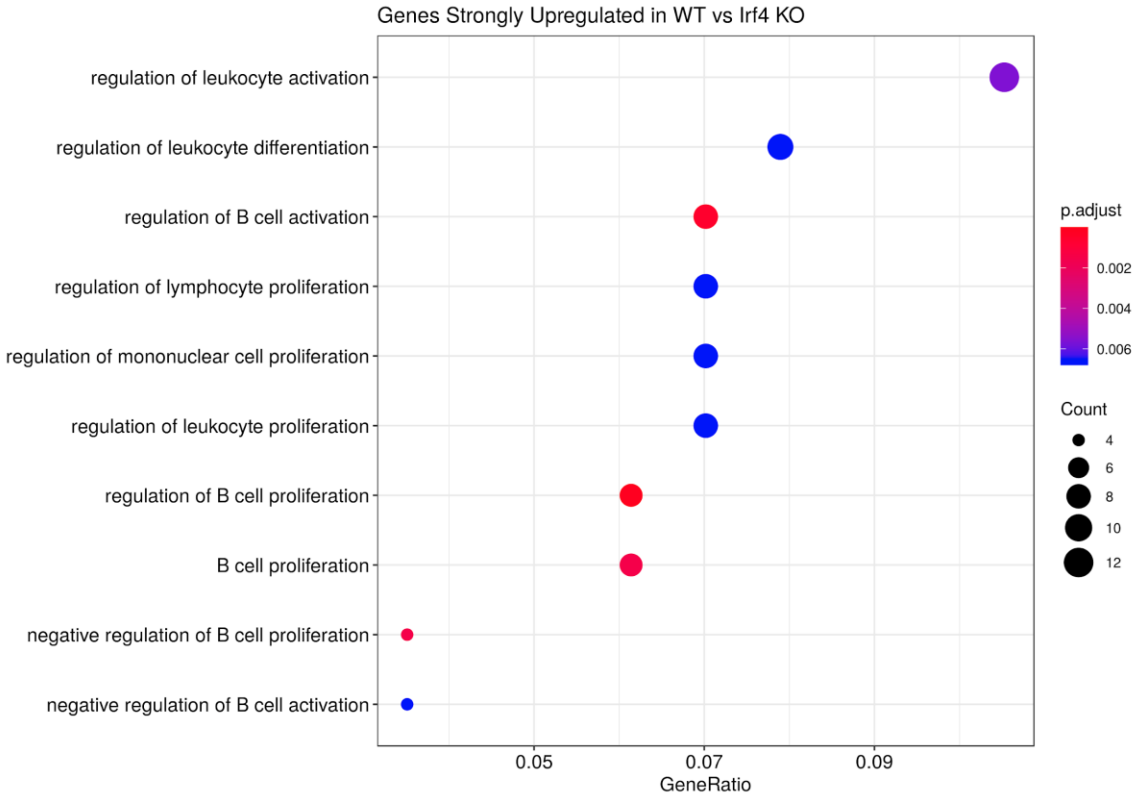
The plot was created using Gviz (Hahne and Ivanek 2016). Fragment pileup was normalized by the number of mapped fragments in each replicate.

Figure 1J - Knockout of *Irf4* Changes Transcriptome



Legend: The volcano plot on the left indicates genes that were differentially expressed in the "WT Flox" (red) and "Irf4 cKO" (green) conditions. Blue genes were not differentially expressed. The barplot indicates the number of differentially expressed genes found in each condition. (Genes with an adjusted pvalue <0.05 are considered differentially expressed.)

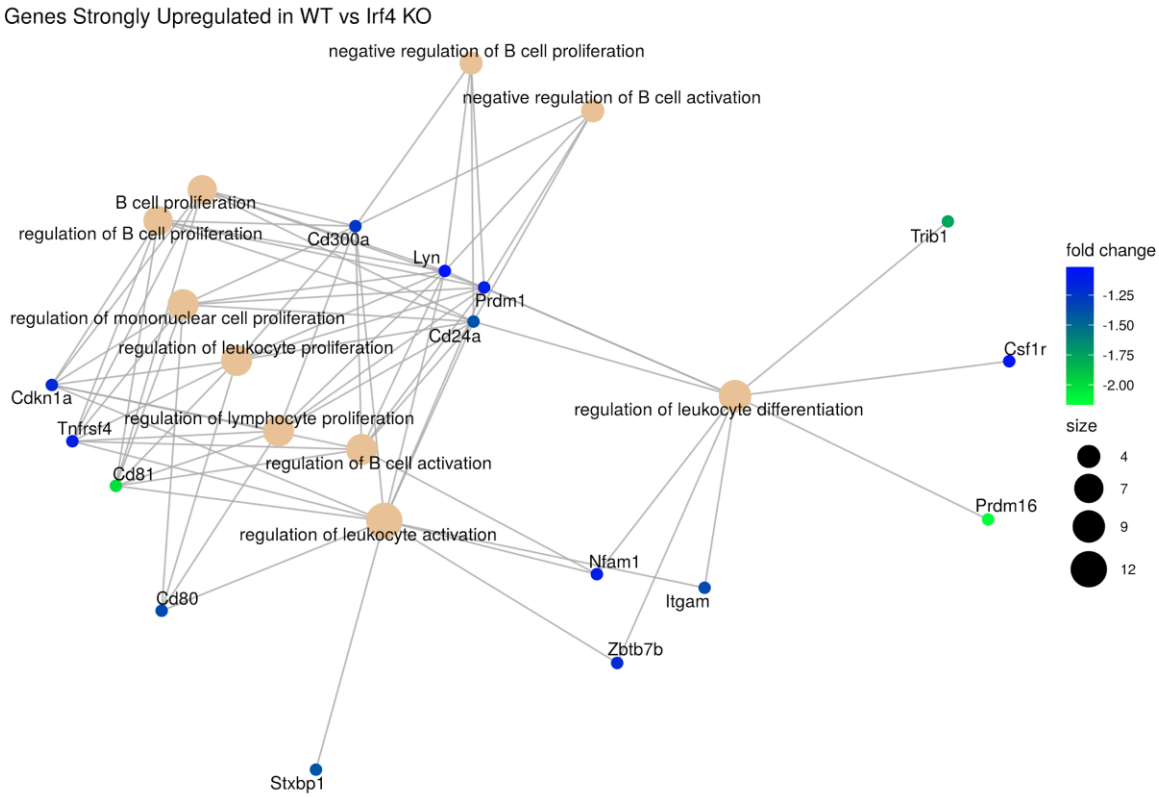
Figure 1K - GO Enrichment for Genes Upregulated in WT vs *Irf4 KO*



Legend: Top-scoring significant GO pathways by hypergeometric test (enrichGO in clusterProfiler package) for genes that were differentially upregulated at least twofold in the WT vs *Irf4 KO* test (n=127). All pathways belong to the Biological Process category. "Gene ratio" represents the percentage of genes in the "differentially up" set that were also in the GO pathway.

The GO enrichment test and plot were carried out with the "clusterProfiler" R package. (Yu et al 2012)

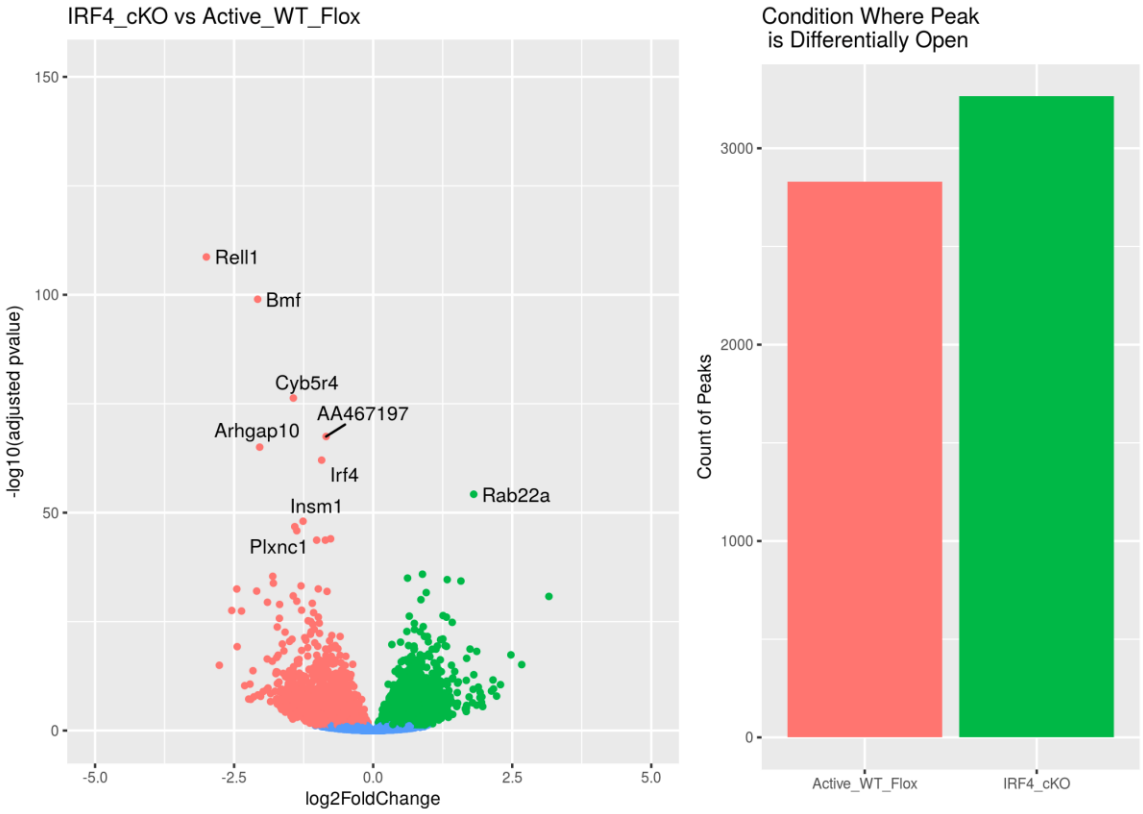
Figure 1L - Several Genes Lost in the *Irf4 KO* Condition are Shared Across Key Immune Pathways



Legend: Each beige point represents one of the top ten GO enrichment categories (ranked by pvalue) that were strongly upregulated (fold change >=2) in the WT vs *Irf4* KO condition.

The GO enrichment test and plot were carried out with the “clusterProfiler” R package. (Yu et al 2012) Note: “fold change” here is the log2 Fold Change estimate from DESeq2.

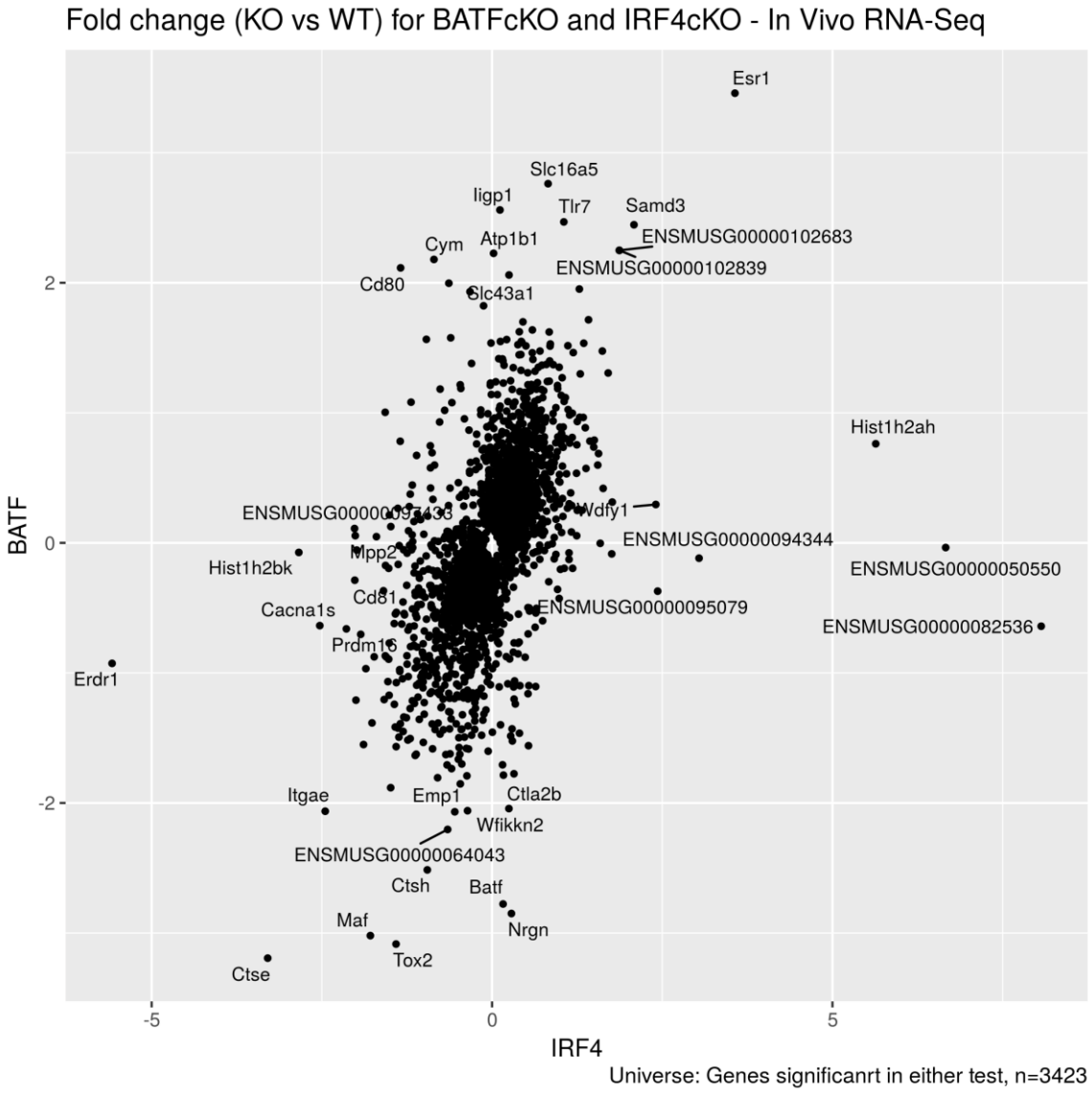
Figure 1M - Knockout of *Irf4* Remodels Chromatin Landscape



Legend: The volcano plot on the left indicates regions on the genome that were differentially accessible in the "WT Flox" (red) or "Irf4 cKO" (green) conditions. Blue loci were not differentially accessible. The barplot indicates the number of differentially accessible regions found in each condition. (Regions with an adjusted pvalue <0.05 in the test are considered differentially accessible.)

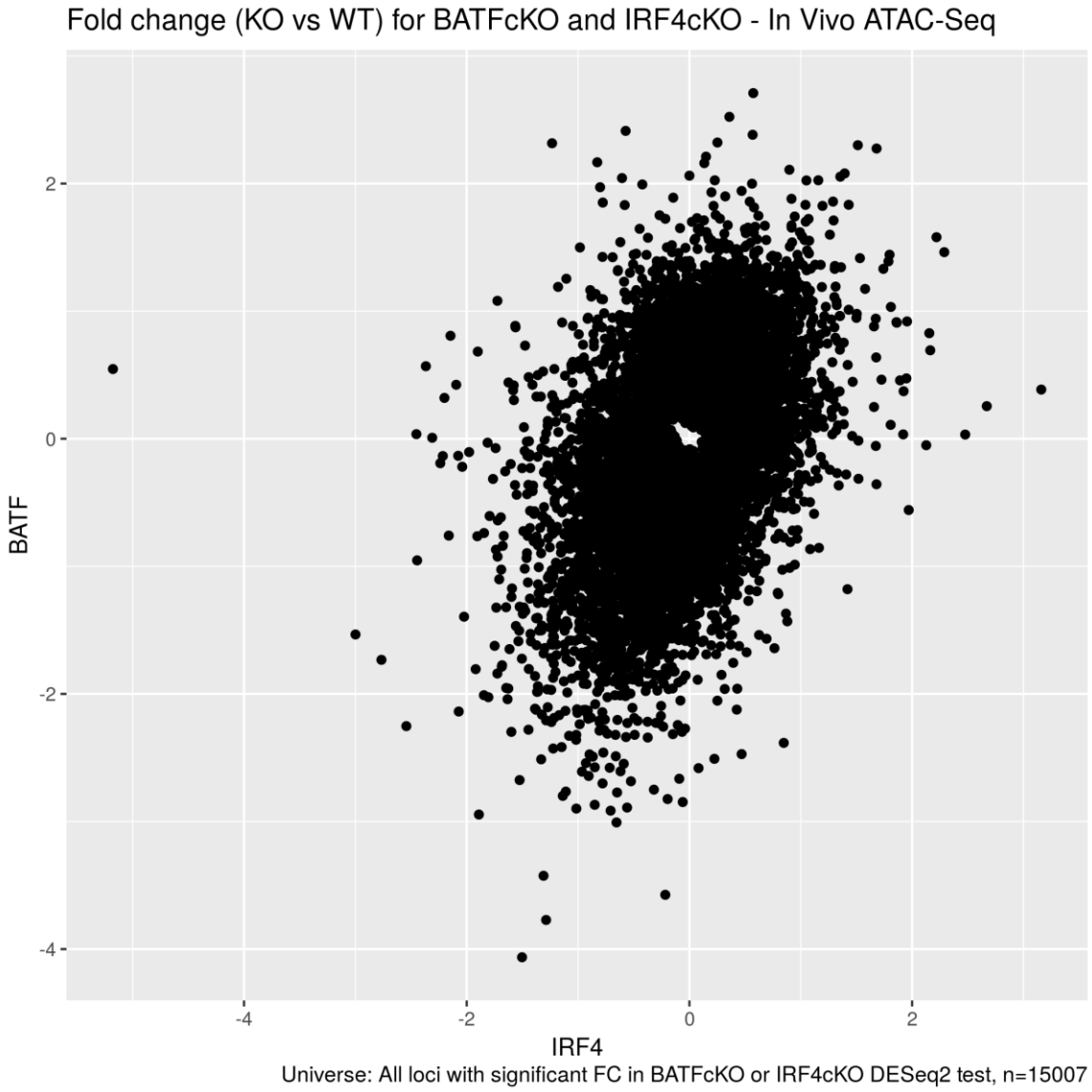
GREAT (McLean et al 2010) was used to map loci to genes they may regulate, and the closest gene for the most differential loci are displayed.

Figure 1N - *Batf* and *Irf4* KO Display Similar Effects for Many Genes



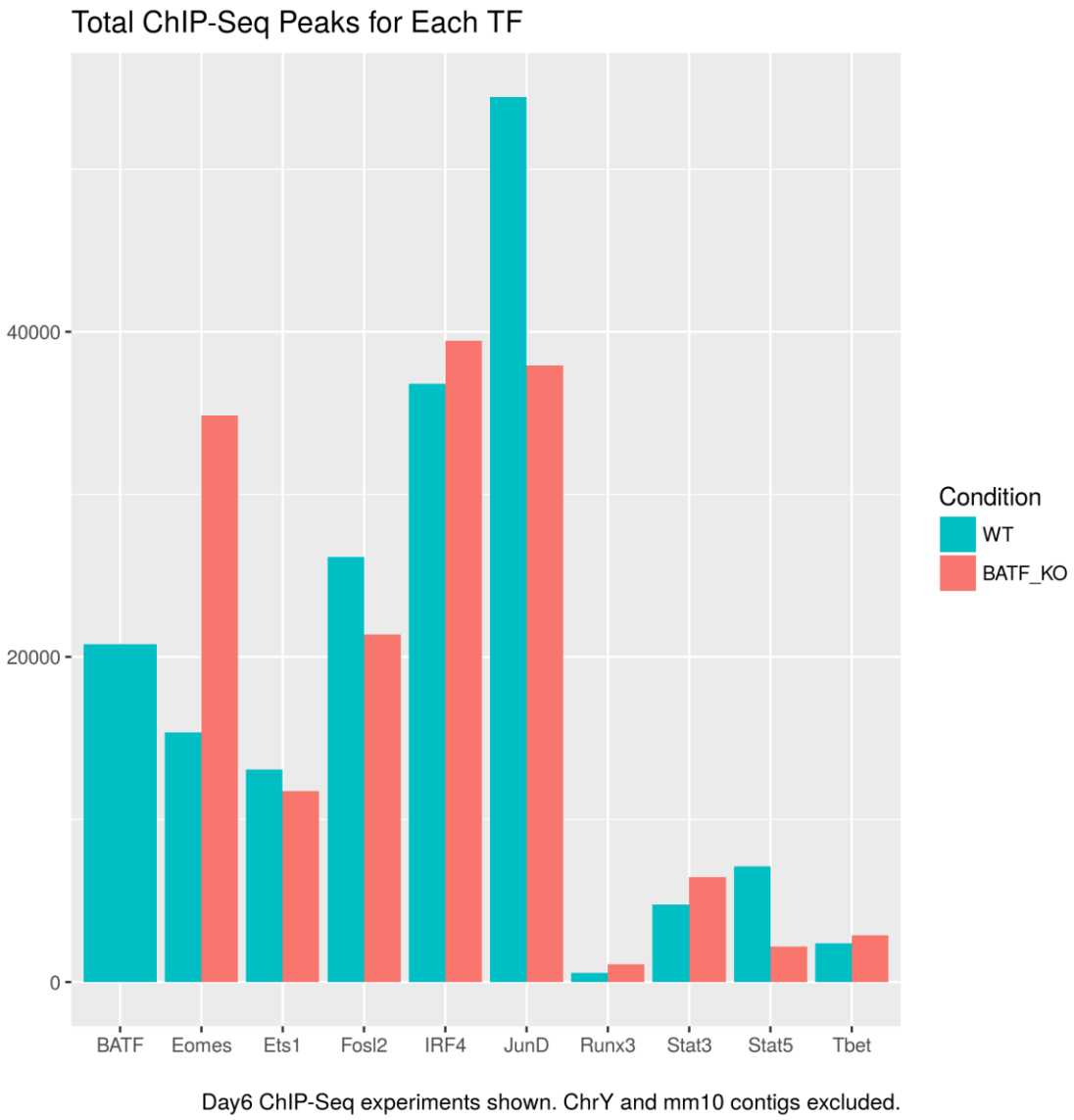
Legend: The plot above shows results for log2(KO/WT) gene expression estimated by DESeq2 for the BATFKO vs WT and IRF4KO vs WT samples from the Cre-Lox experiment.

Figure 1O - Batf and Irf4 KO Display Similar Effects for Many ChAR's



Legend: The plot above shows results for log2(KO/WT) ATAC-Seq Tn5 counts estimated by DESeq2 for the BATFKO vs WT and IRF4KO vs WT samples from the Cre-Lox experiment.

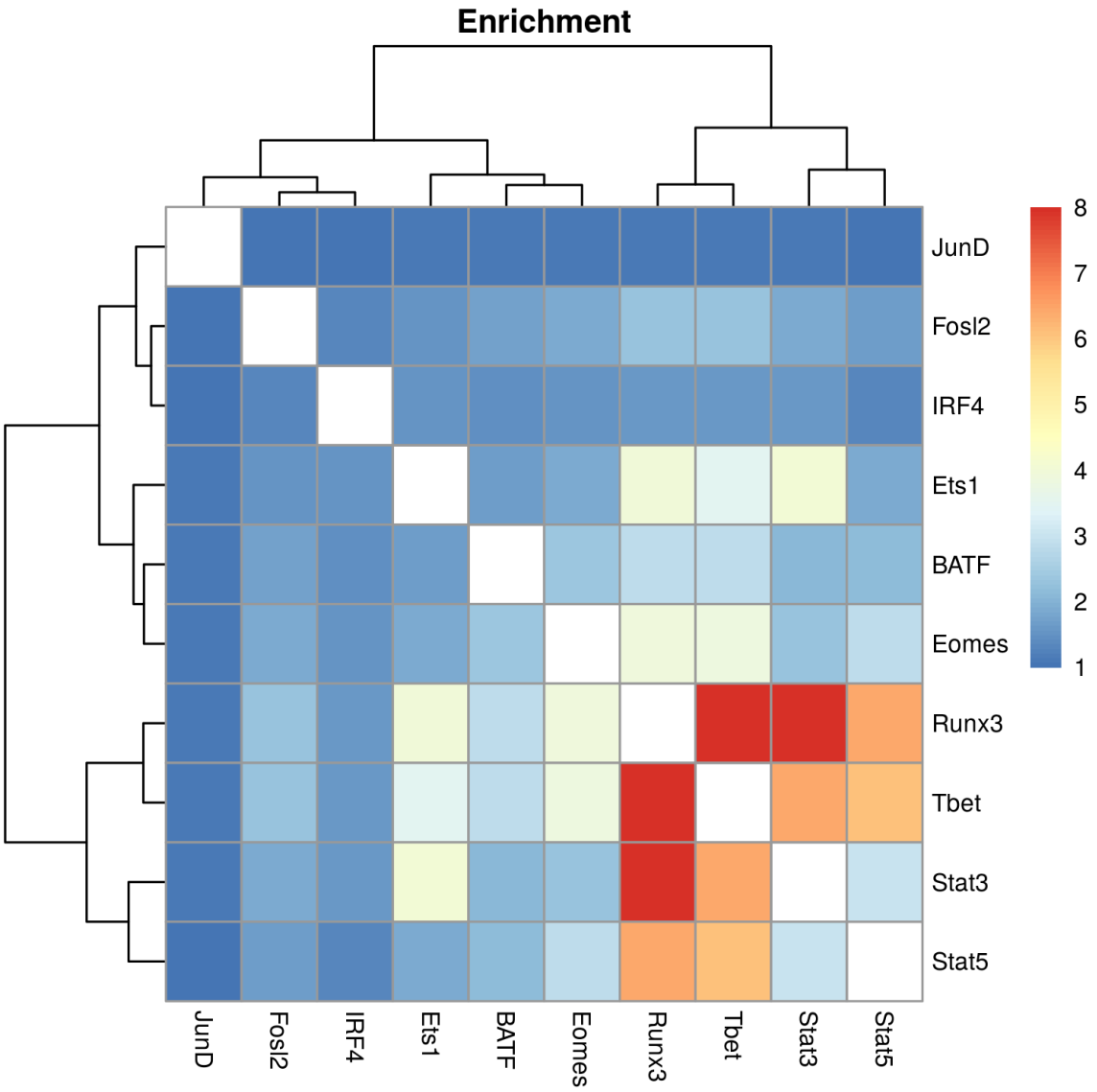
Fig 2A - Total ChIP-Seq Peaks for Each Transcription Factor



Legend: The plot above displays the count of peaks called.

Each combination of transcription factor and condition had two replicates, with the exception of Runx3-WT, which had three replicates. ChIP-Seq peaks were called in each replicate, versus a control sample, using macs2 with a qvalue cutoff of 0.05 and overlapping peaks among replicates were merged. A region was considered a valid peak for a transcription factor if it had a qvalue below 0.05 in at least two replicates, and a qvalue below 0.001 in at least one of the replicates.

Fig 2B - *Runx3* and *Tbet* are enriched for co-binding with other key transcription factors

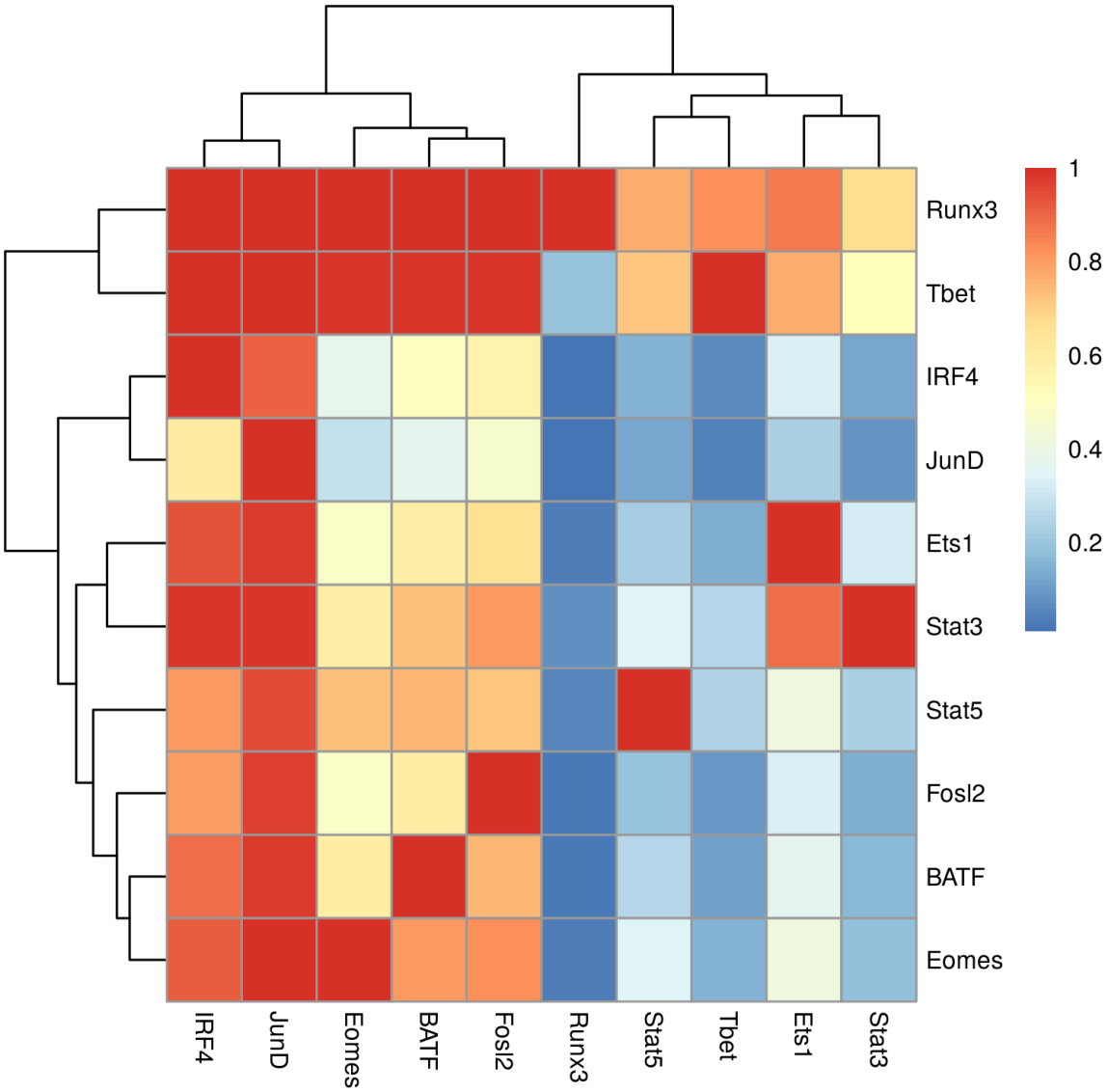


Legend: The above plot shows fold enrichment for binding of one transcription factor within the loci of another. Rows and columns were clustered using Euclidean distance in the pheatmap package.

Fold enrichment (for TF A in TF B) = (Percent of TF A Peaks Co-Bound By TF B)/ (Percent of All Day6 WT ChIP-Seq Peaks Bound by TF B).

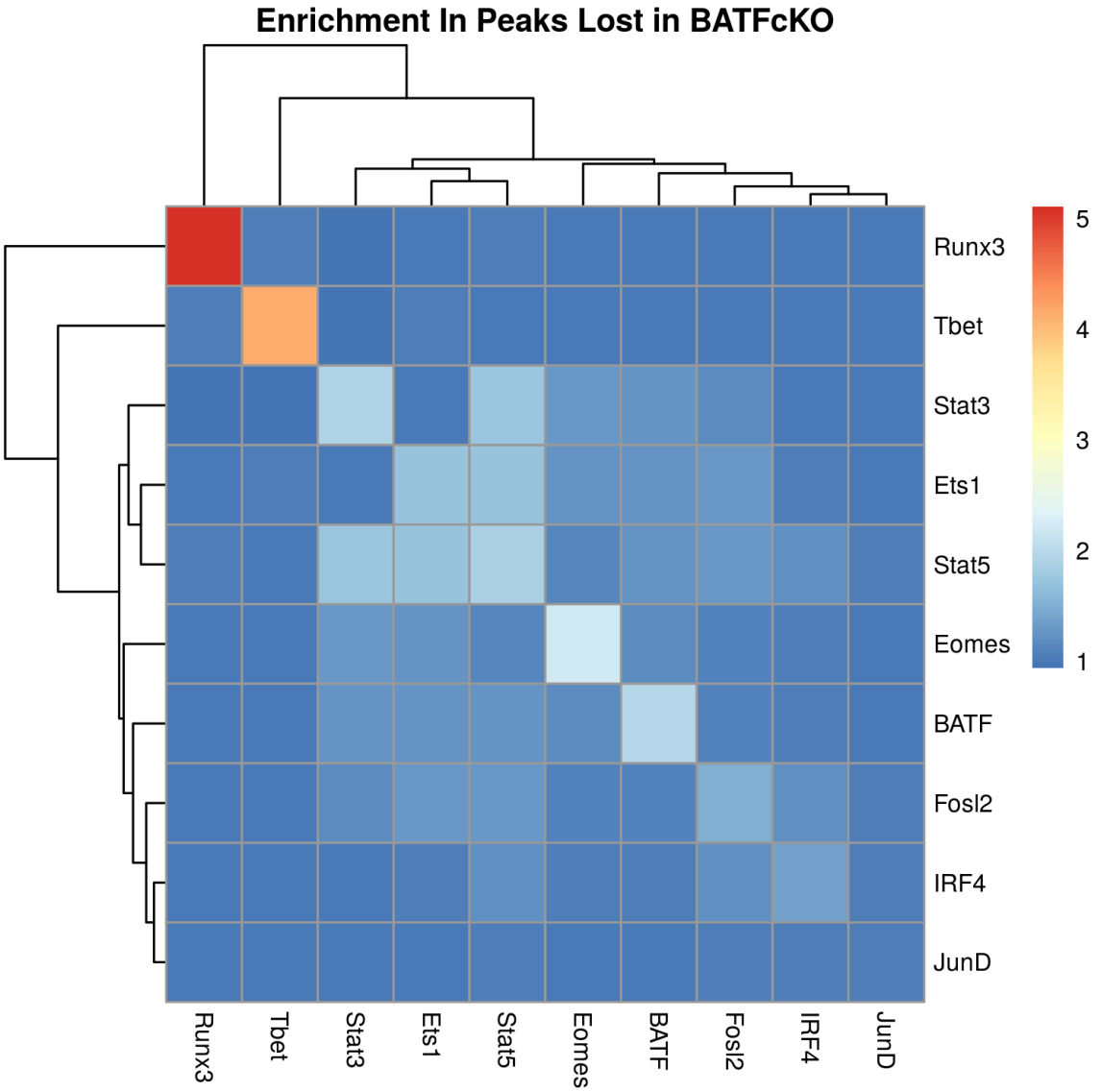
Note: Runx3 fold enrichment for TBet binding is 20.7, and Runx3 fold enrichment for Stat3 was 8.5. These values were capped at 8 to avoid distorting the plot.

Figure 2C - Percent of Row TF Peaks Co-Bound by Column TF



Legend: Each cell contains the percentage of row transcription factor peaks co-bound by the column transcription factor. Rows and columns were clustered using Euclidean distance in the pheatmap package.

Figure 2D - Fold enrichment of transcription factor binding within peaks that are less accessible when *Batf* is knocked out

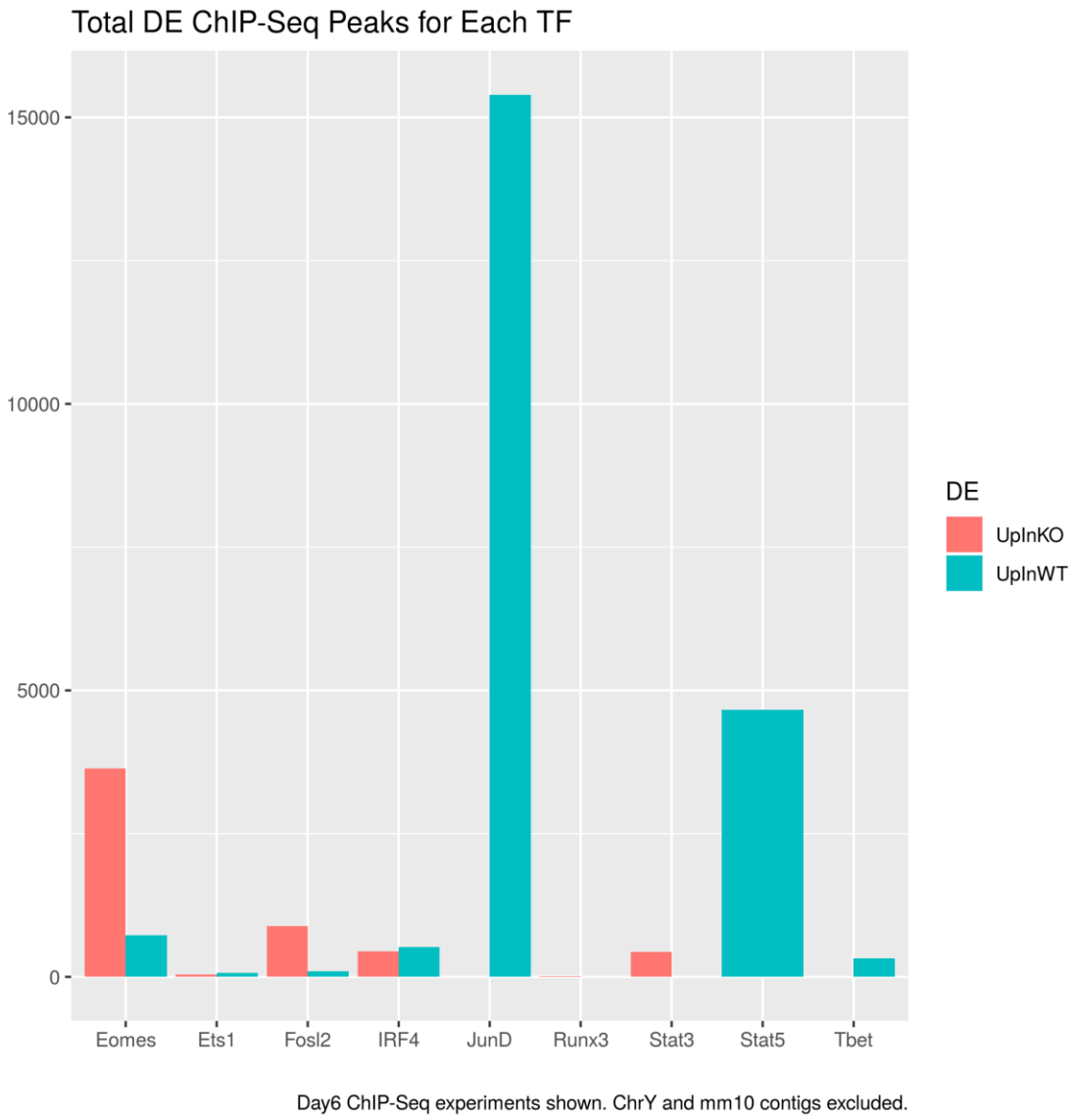


Legend: Peaks lost in the *Batf* KO were identified has having log2(KO/WT) significantly higher than 0, using DESeq2 (adjusted pvalue<0.05).

The diagonal cells display fold enrichment for TF Binding in the "lost peaks" = (Percent of "TF A" Peaks lost in Batf KO) / (Pct of Day 6 WT ChIP Peaks lost in BATF KO).

Cells on the off-diagonal contain fold enrichment for co-binding in the peaks lost in the *Batf KO*. This is calculated as [Percent of Loci Bound by TF's A and B Lost in Batf KO / max(Pct of TF A Peaks Lost in Batf KO, Pct of TF B Peaks Lost in Batf KO)]. The heatmap was created with the pheatmap package.

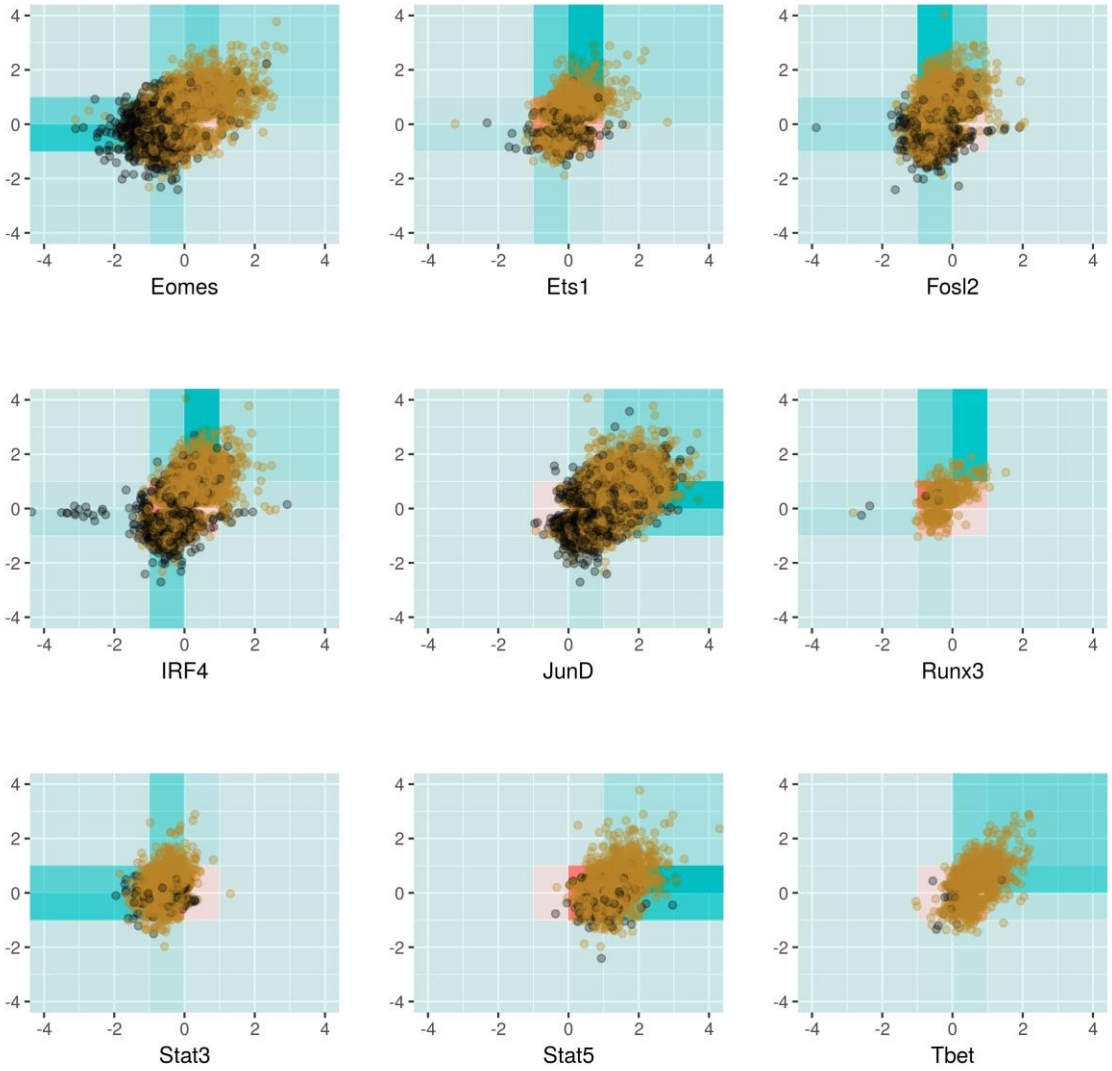
Figure 2E - Count of Differentially Bound ChIP-Seq TFs for WT vs *Batf* KO



Legend: The plot displays the count of peaks for each transcription factor identified as differentially bound in the WT or Batf KO samples. Differentially bound ChIP-Seq peaks were identified using the csaw package (Lun and Smyth 2016) .

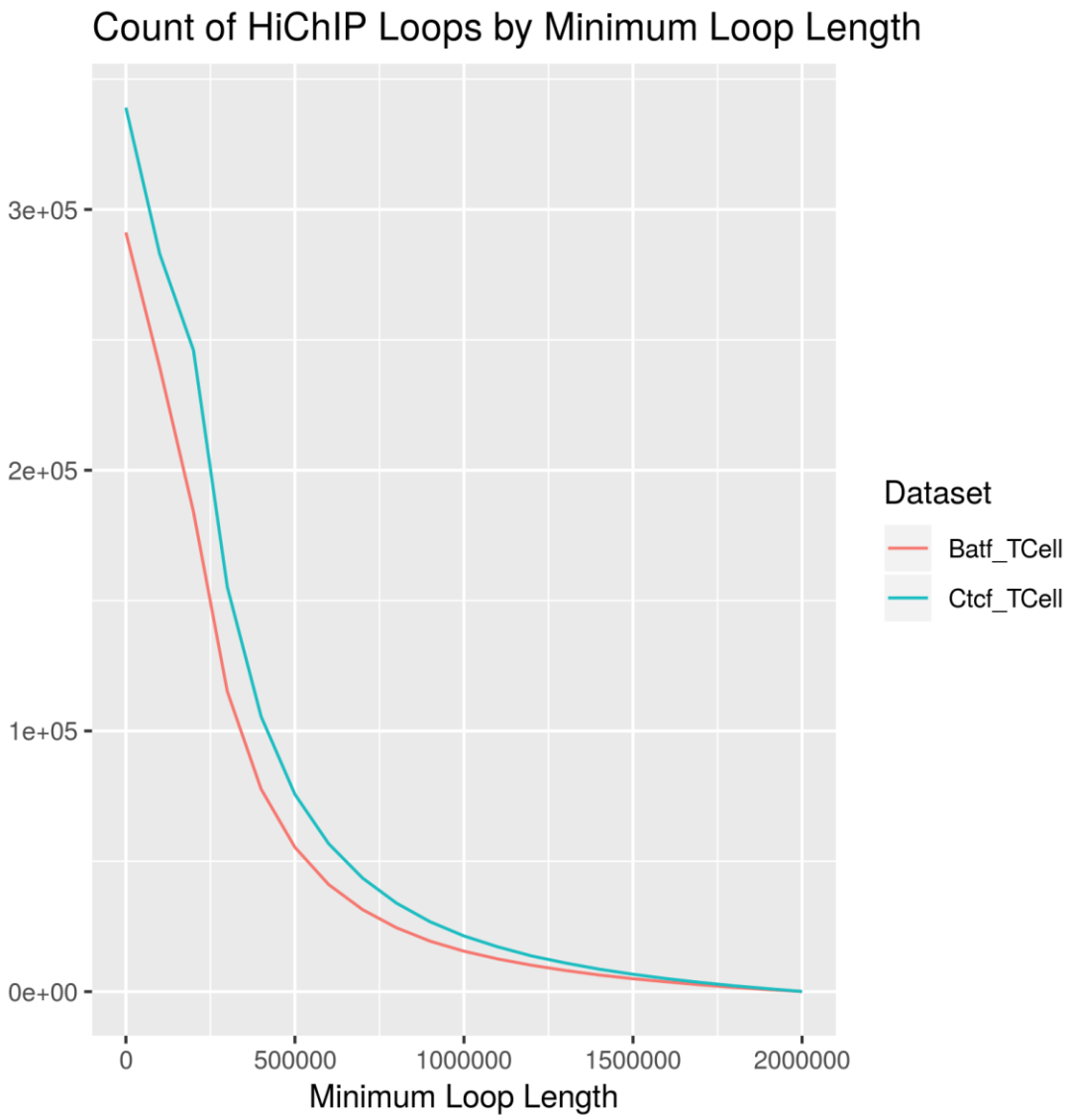
Differential analysis was only carried out on regions where a peak was identified as bound by a transcription factor in either the WT or Batf KO. Regions were considered differentially bound if csaw estimated a log2 fold change greater than one, and a qvalue <0.05.

Figure 2F - *Batf* knockout affects transcription factors through changes in chromatin accessibility and co-binding affinity.



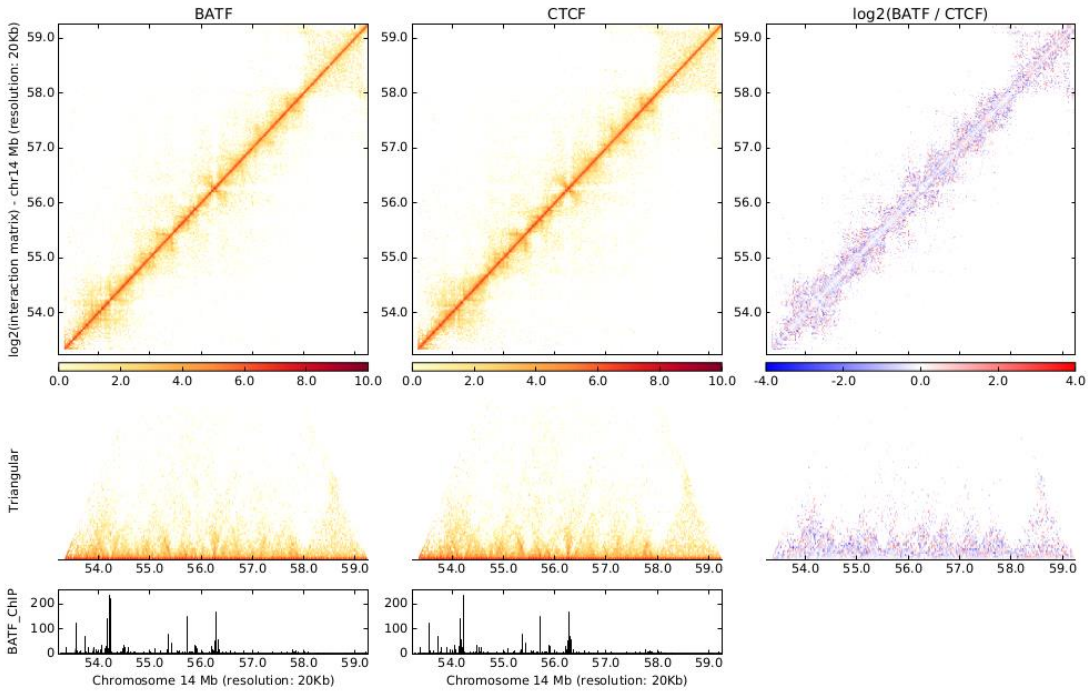
Legend: The universe for each plot consists of all loci that are either 1) differentially bound by the transcription factor in the ChIP data for WT vs *Batf KO* with a fold change greater than 2, or 2) are differentially accessible in the ATAC-Seq data.
The y-axis displays the change in chromatin accessibility log2(WT/BatfKO) and the x-axis displays the change in transcription factor binding log2(WT/BatfKO). Gold points are bound by Batf in the Day6 WT data.

Figure 3A - Count of HiChIP Loops by Minimum Loop Length



Legend: This plot displays the count of HiChIP loops (long range chromatin interactions) identified when using different values for minimum loop length.

Figure 3B - HiChIP Data for T Cells, using *Batf* and *Ctcf* as the ChIP targets.



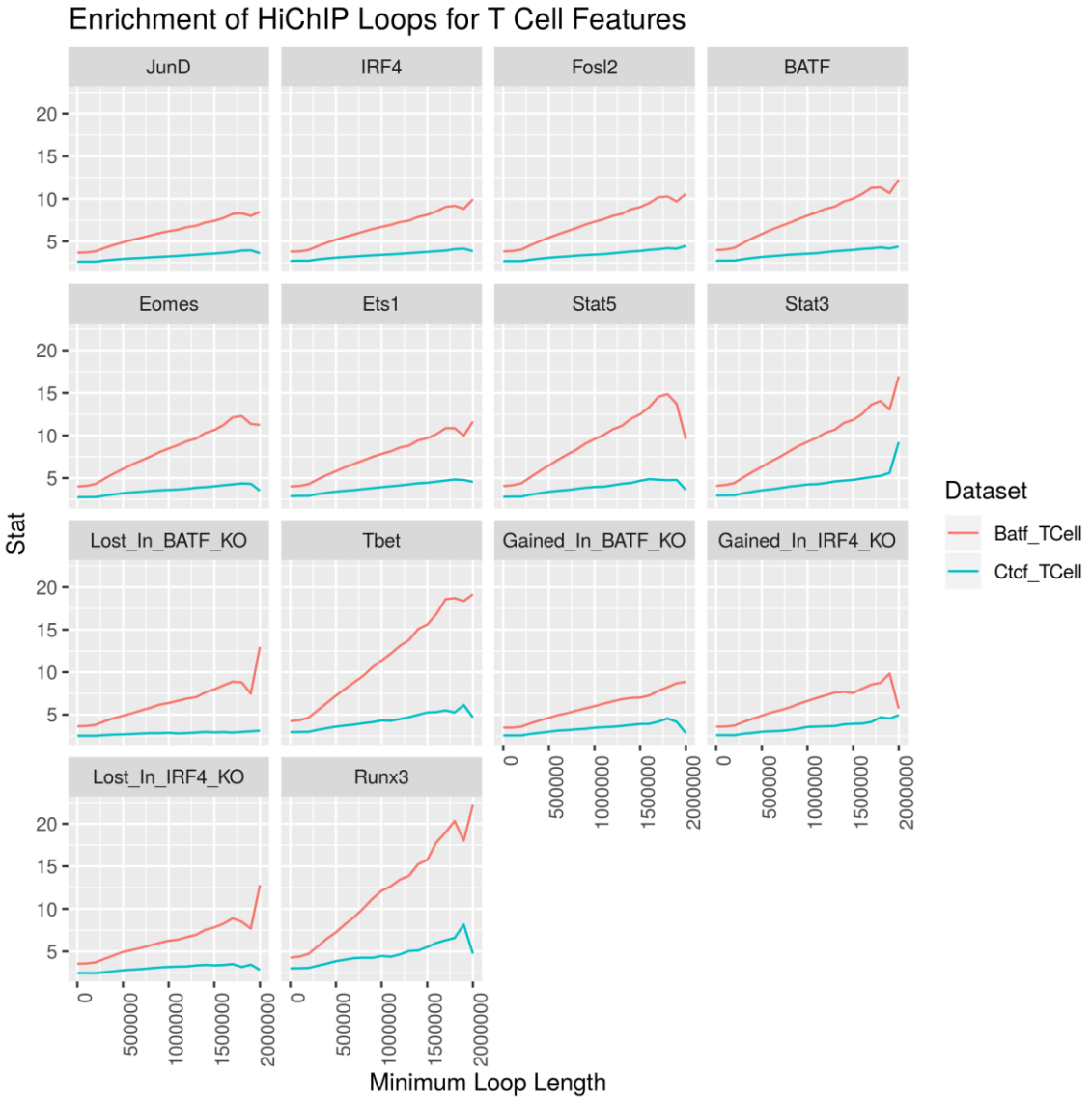
Legend: The first row of plots illustrate the normalized counts of contacts from obtained from HiC-Pro for the Batf and Ctcf T Cell samples surrounding Gzmb on Chromosome 14. The final plot represents the log2 ratio of these normalized counts. To create these plots, the genome is binned into regions of 20 Kbp.

The middle row displays the same data, along the diagonal of the heatmap.

The bottom row displays bedgraph data for Batf binding from the 10ug Batf Day 6 ChIP WT sample.

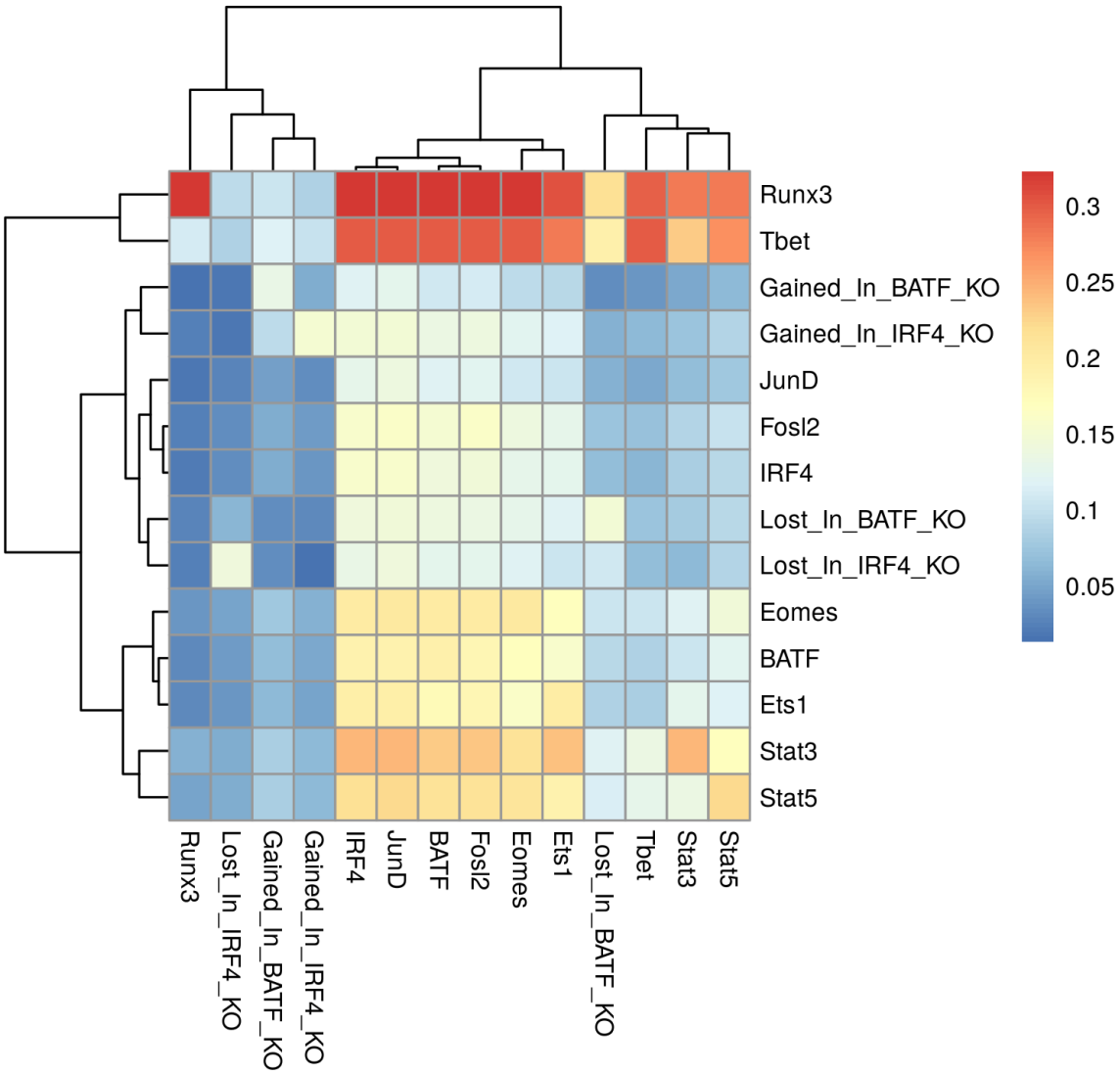
The plot was created using the HiC Plotter package for Python.

Figure 3C - HiChIP Loop Anchors are Enriched for Binding by Key Transcription Factors and Differentially Accessible ChAR's.



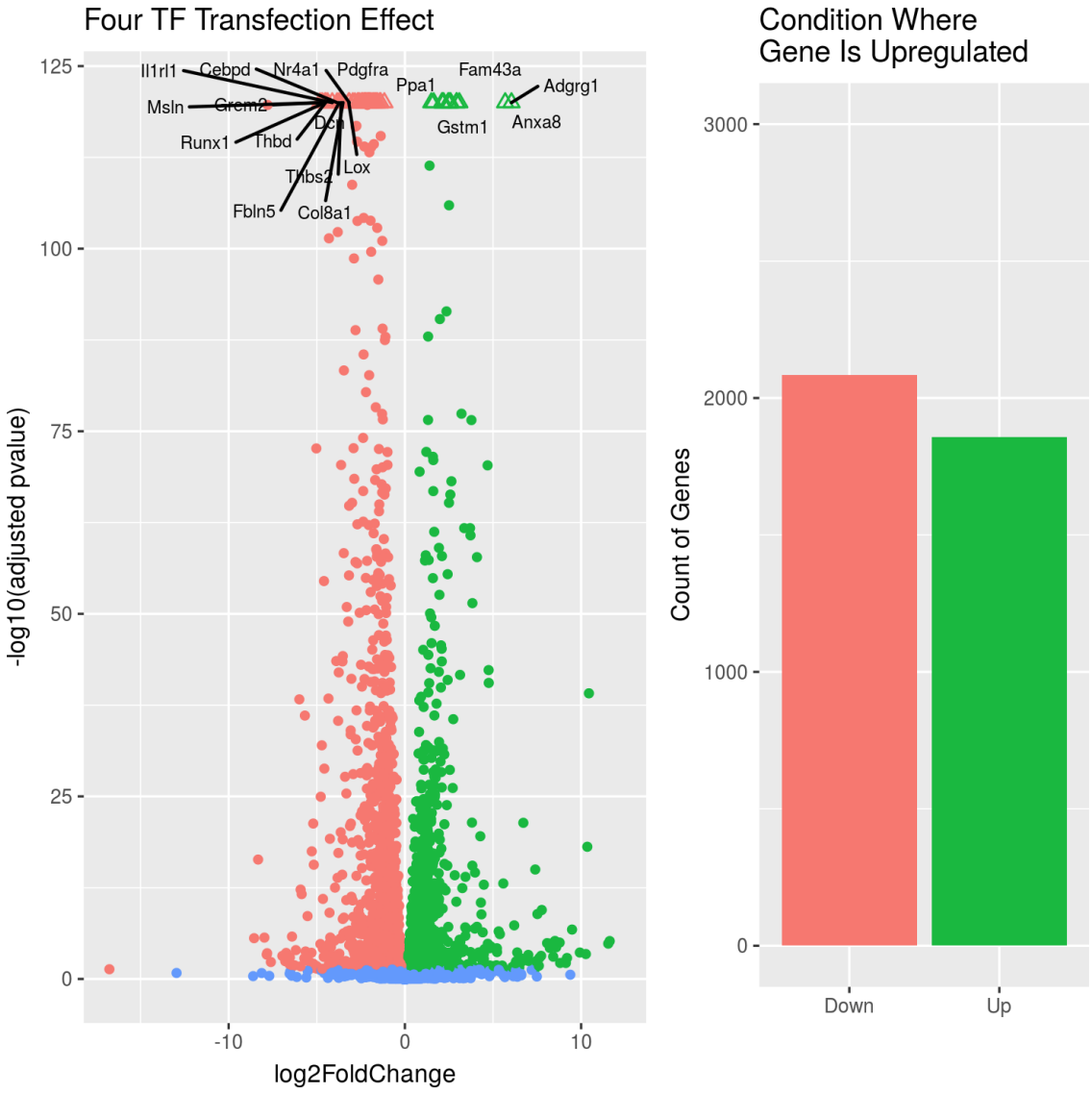
Legend: Each plot displays the fold enrichment of HiChIP loop anchors for differential ChAR's and T Cell Transcription factors. The cutoff for minimum loop length varies along the x-axis.

Figure 3D - High percentages of *Runx3* and *Tbet* Peaks Participate in Long Range Contacts with other Regulatory Features



Legend: Each cell contains the percentage of peaks for a given feature that are connected by a HiChIP loop length of >= 1,000,000 bp to a column feature. The plot was created with the pheatmap package.

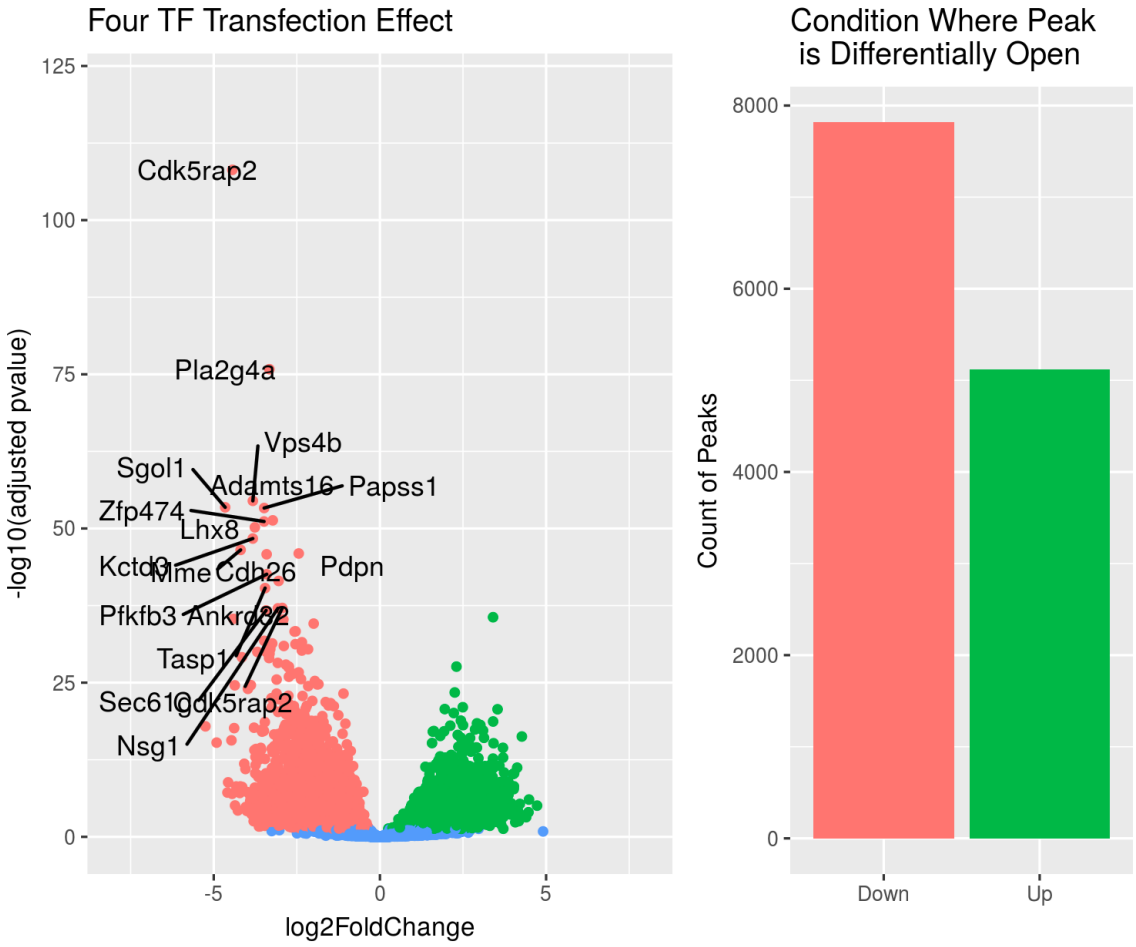
Figure 4A - Expression of Four Key CD8+ T Cell TFs in Fibroblasts Changes Transcriptome



Legend: The volcano plot on the left indicates genes that were differentially expressed in the Fibroblast samples without transfection and Doxocycline treatment (red) and the fibroblast samples with transfection and Doxocyline treatment (green) conditions. Blue genes were not differentially expressed. The model used in DESeq2 was GeneExpression ~ DoxTreatment + FourTF_Transfection + DoxTreatment * FourTF_Transfection, and the results above are with respect to the "DoxTreatment * FourTF_Transfection" coefficient. This model was used to control for effects that the Doxocycline treatment, or transfection without activation, may have on the fibroblasts.

The barplot indicates the number of differentially expressed genes found in each condition. (Genes with an adjusted pvalue <0.05 are considered differentially expressed.)

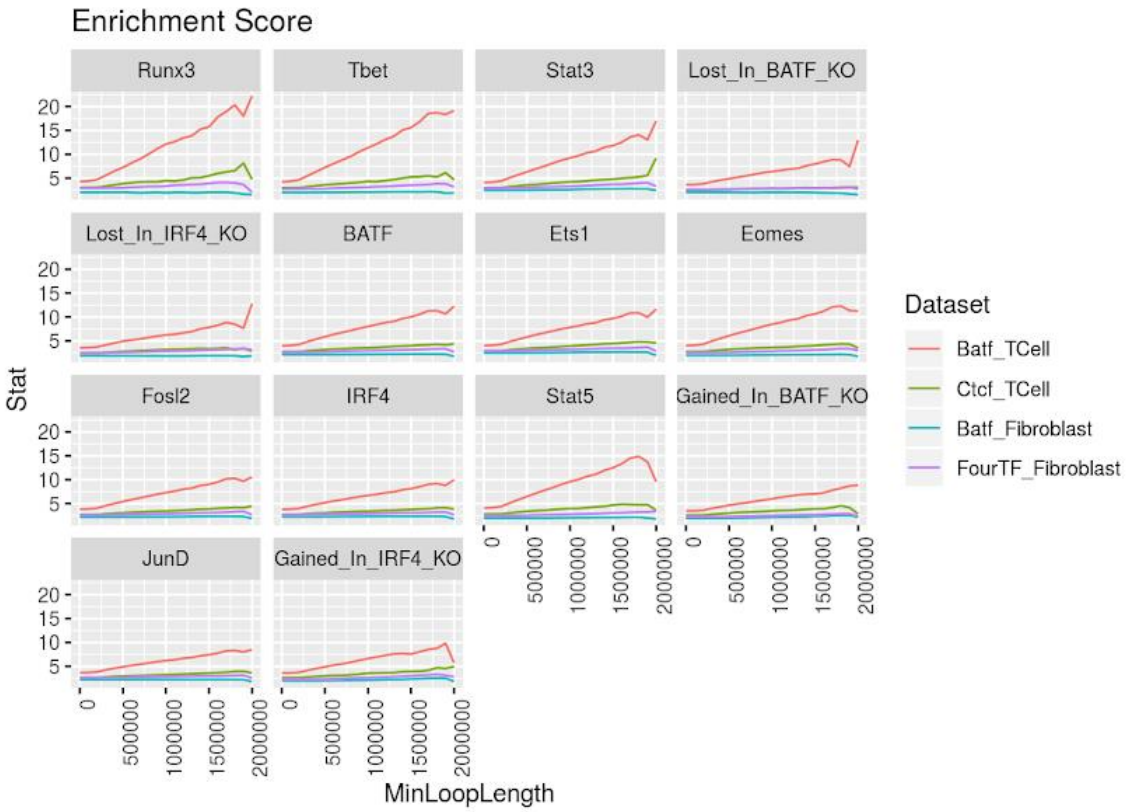
Figure 4B - Expression of Four Key CD8+ T cell TFs in Fibroblasts Remodels the Chromatin Landscape



Legend: The volcano plot on the left indicates loci that were differentially expressed in the Fibroblast samples without transfection and Dox treatment (red) and the fibroblast samples with transfection and Dox treatment (green) conditions. Blue loci were not differentially expressed. The model used in DESeq2 was Tn5Cuts ~ DoxTreatment + FourTF_Transfection + DoxTreatment * FourTF_Transfection, and the results above are with respect to the "DoxTreatment * FourTF_Transfection" coefficient. This model was used to control for effects that the Doxocycline treatment, or transfection without activation, may have on the firboblasts.

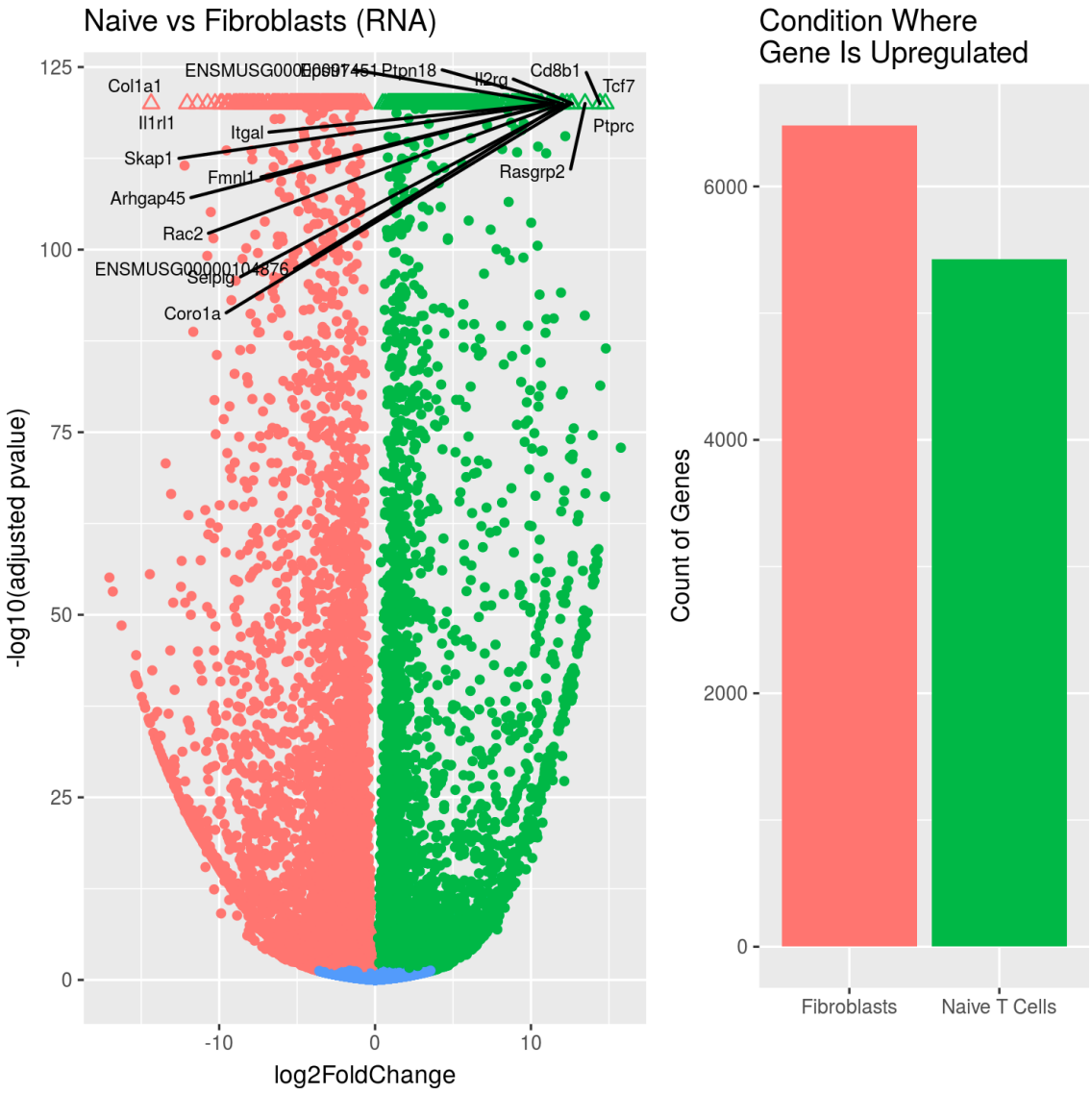
The barplot indicates the number of differentially expressed genes found in each condition. (Loci with an adjusted pvalue <0.05 are considered differentially expressed.)

Figure 4C - Enrichment for CD8+ T Cell Regulatory Features in HiChIP Anchors



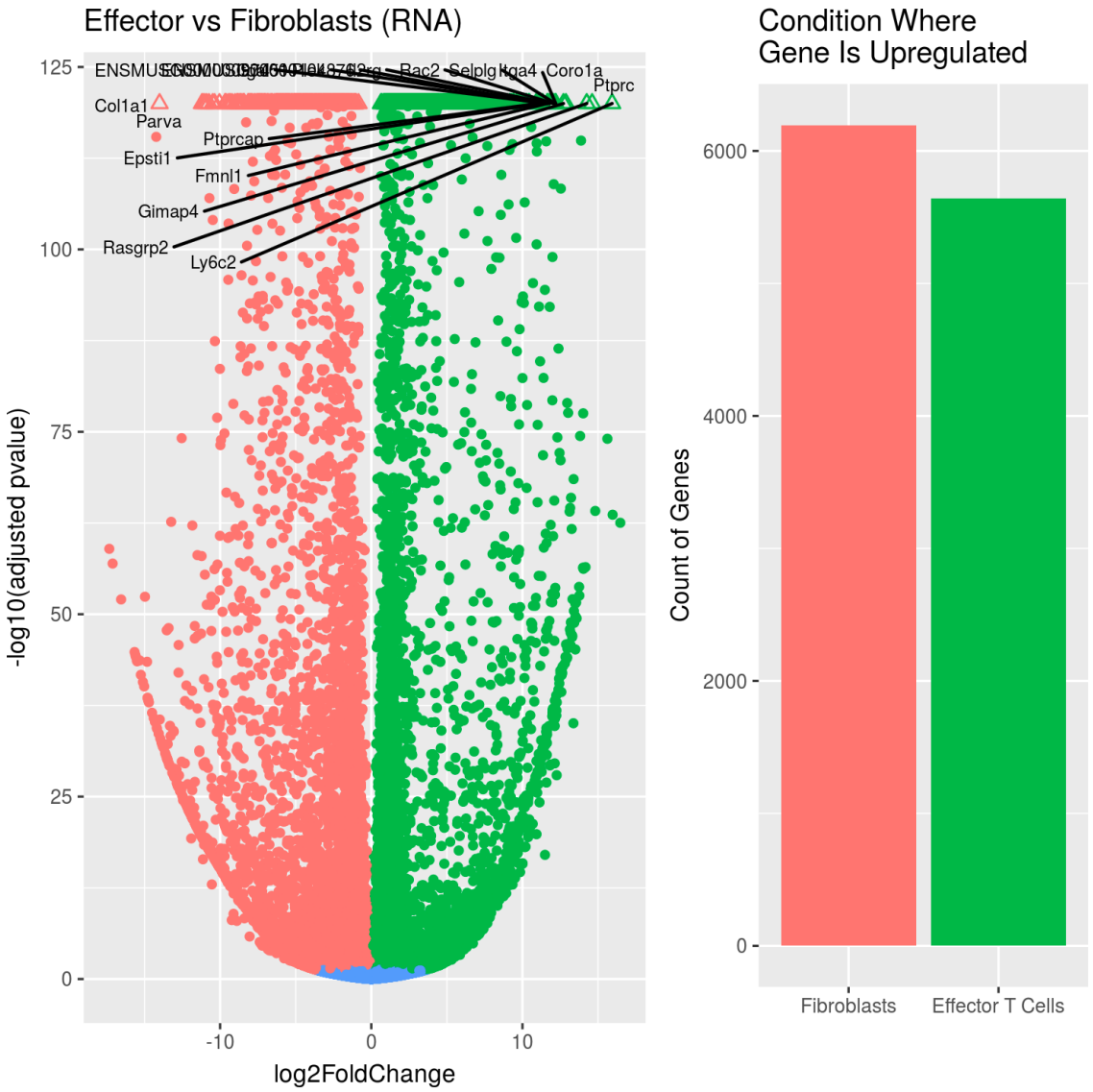
Legend: Each plot displays the fold enrichment of HiChIP loop anchors for differential ChAR's and T Cell Transcription factors. The cutoff for minimum loop length varies along the x-axis.

Figure 5A - RNA-Seq Signature for Naive T Cells



Legend: The volcano plot displays genes that were differentially upregulated in either in Naive T Cells (green) or fibroblasts (red). The top 1,000 genes in each condition where used to create the Naive T Cell signature. Genes upregulated in the Naive T Cell condition were given a value of +1, genes upregulated in the fibroblast condition were given a value of -1.

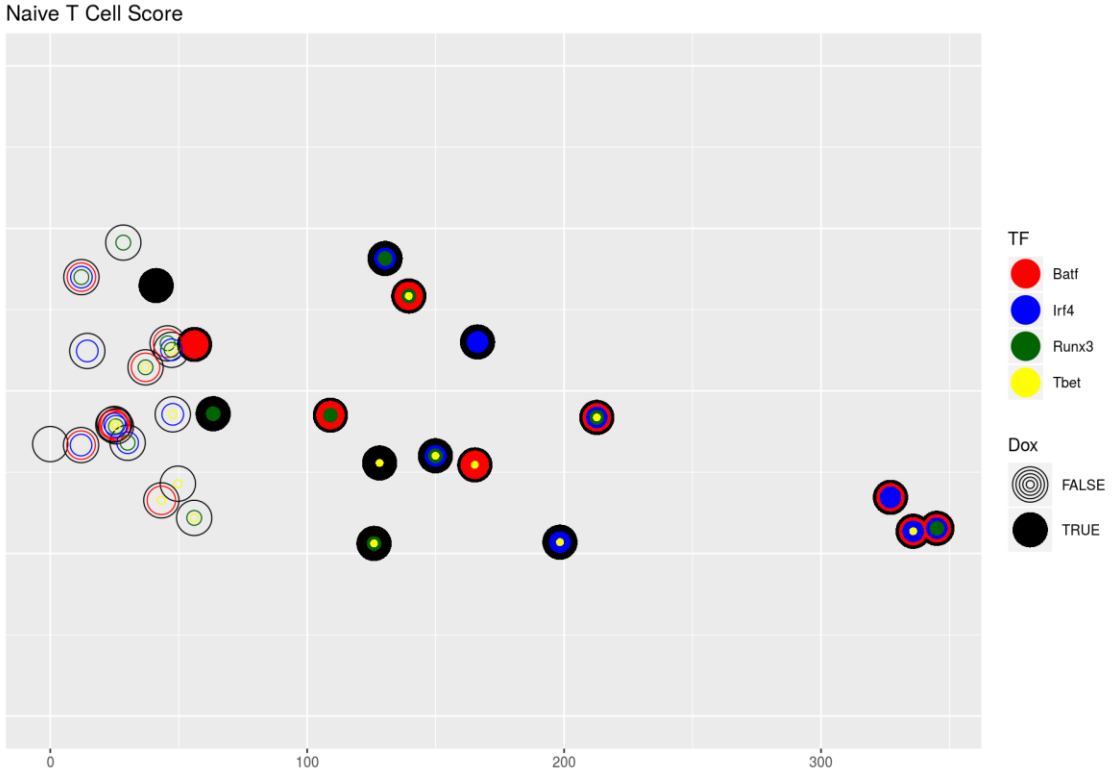
Figure 5B - RNA-Seq Signature for Effector T Cells



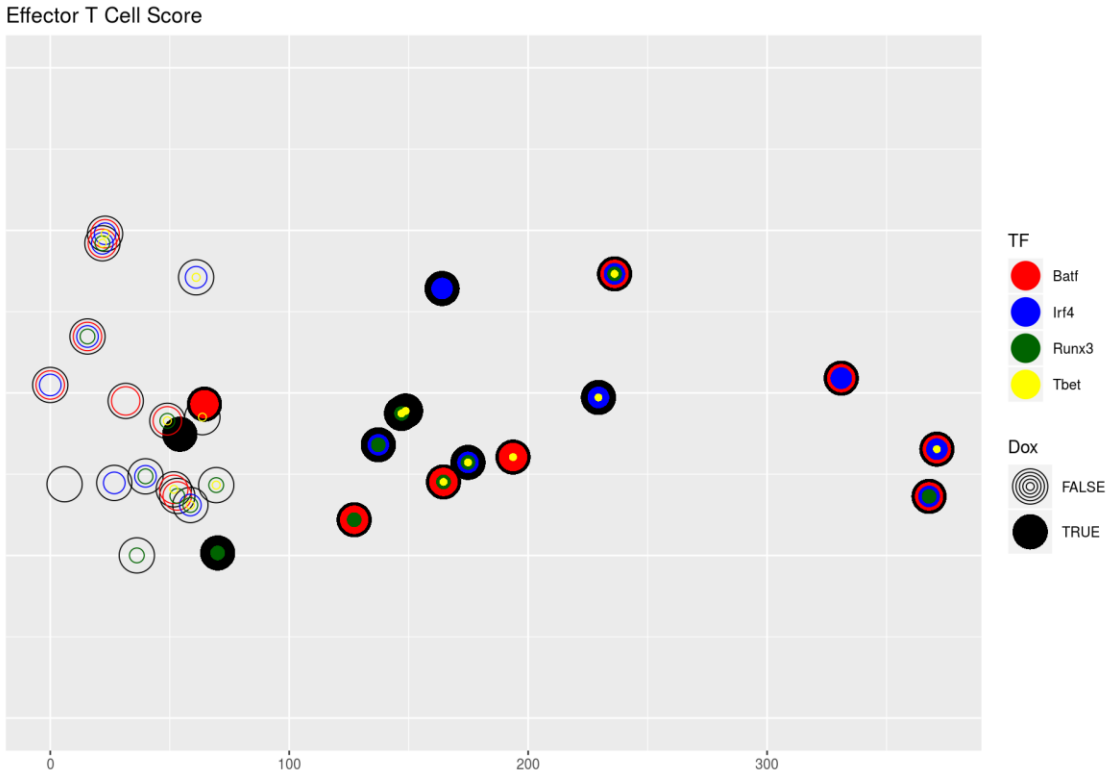
Legend: The volcano plot displays genes that were differentially upregulated in either in Effector T Cells (green) or fibroblasts (red). The top 1,000 genes in each condition where used to create the Effector T Cell signature. Genes upregulated in the Effector T Cells condition were given a value of +1, genes upregulated in the fibroblast condition were given a value of -1.

Figure 5C: T Cell Scores for the Fibroblast Experiments (RNA)

A) Naive T Cell Score

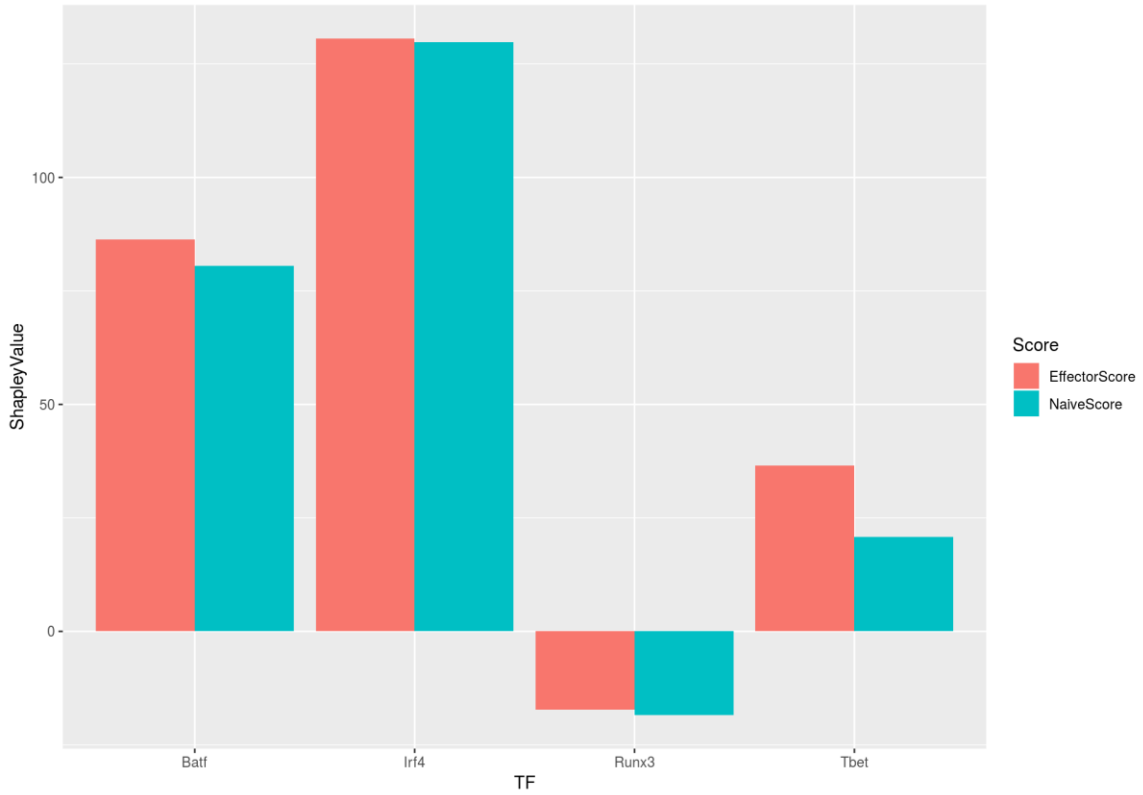


B) Effector T Cell Score



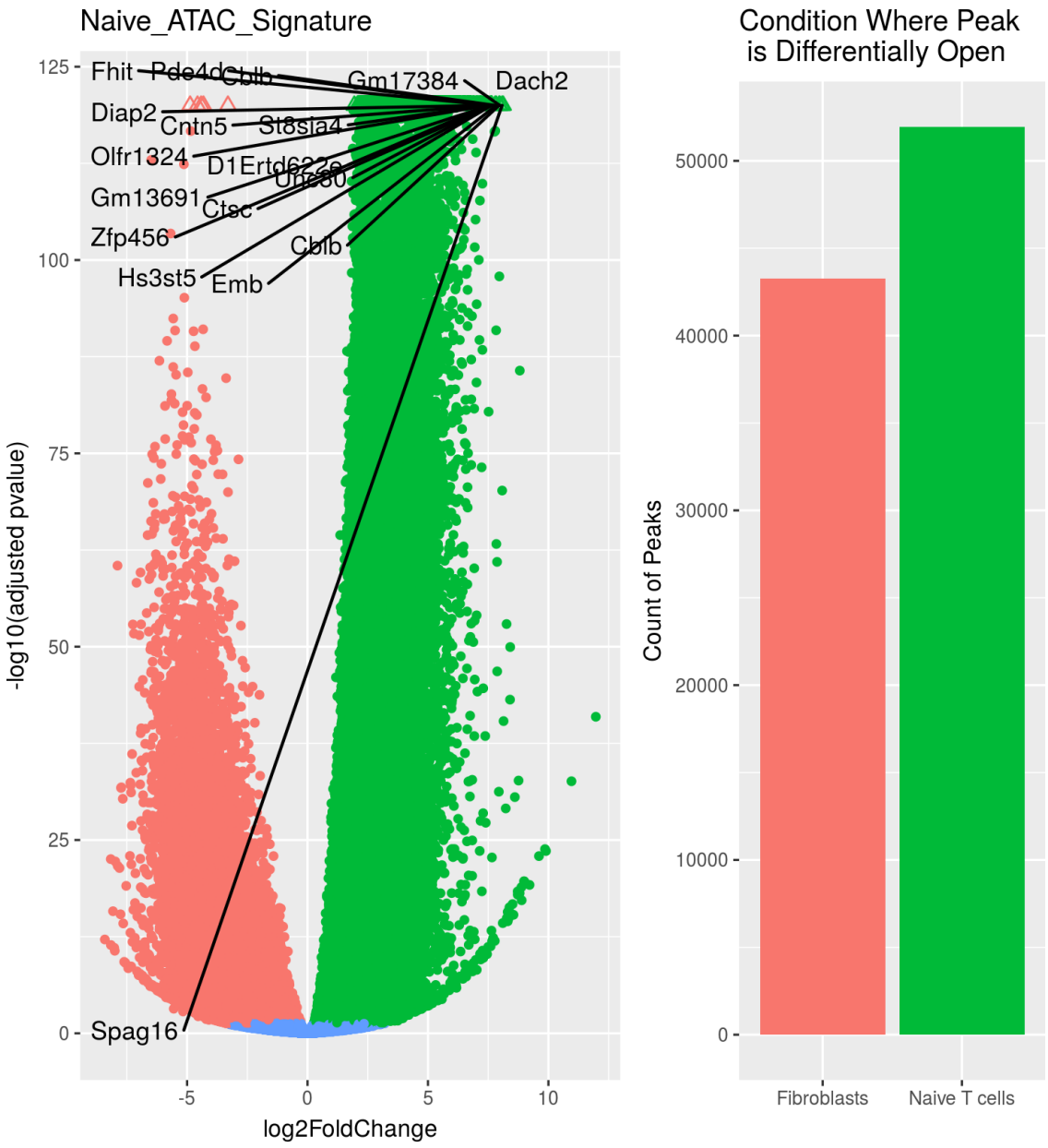
Legend: Each point represents a fibroblast sample. Empty circles represent samples without Doxocycline treatment, solid circles represent Doxocycline treatment. Colors indicate which of the four TFs have been added to the sample. The x-axis represents the T cell signature score. The y-axis is random "jitter" to separate points and make the plot easier to interpret.

Figure 5D - Shapley Value for Each TF for the RNA-Seq T Cell Signatures



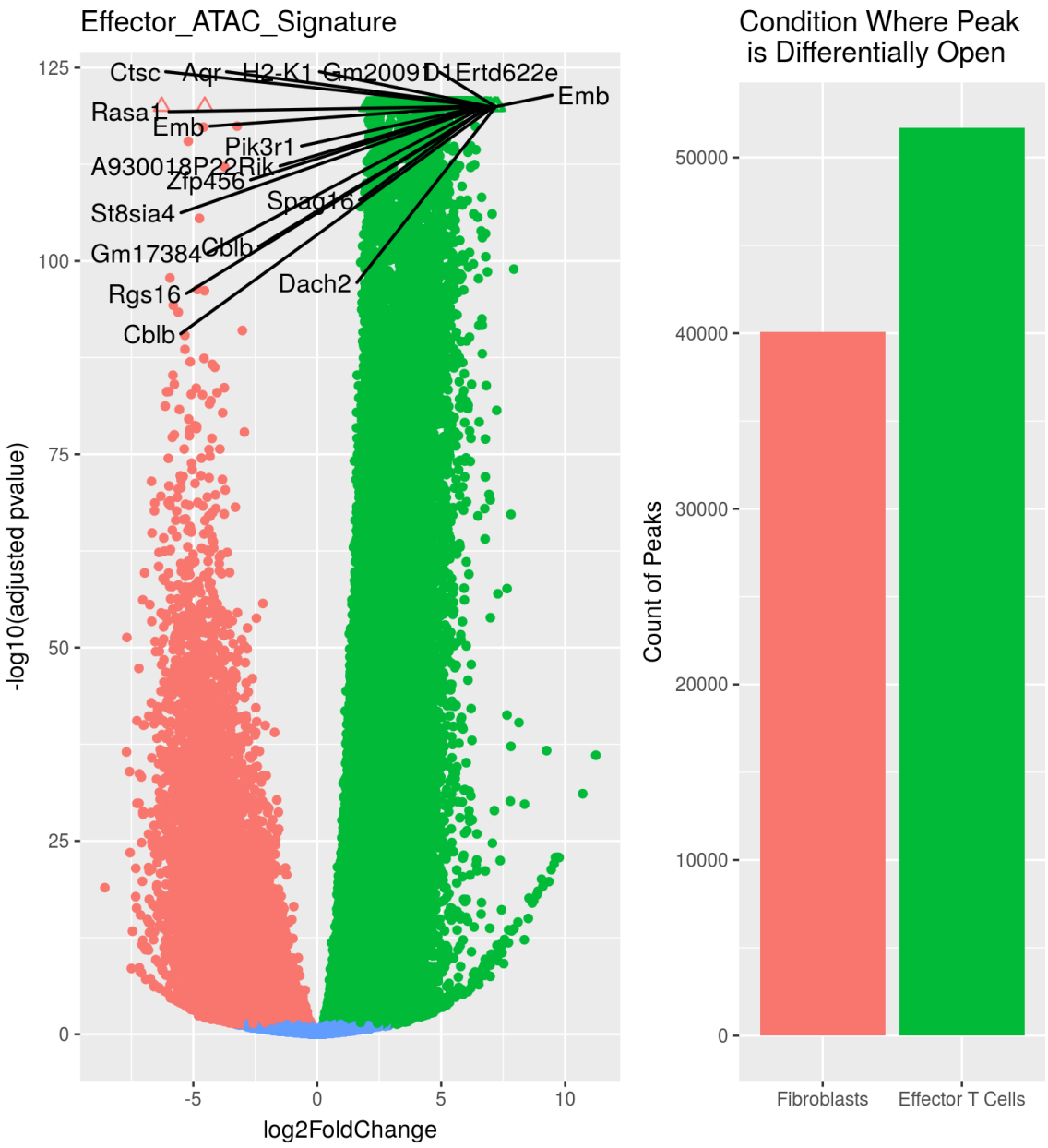
Legend: Each color bar represents the Shapley value, a measure of a TF's contribution to moving fibroblasts towards the given T cell signature.

Figure 5E - ATAC-Seq Signature for Naive T Cells



Legend: The volcano plot displays peaks that were differentially open in either in Naive T cells (green) or fibroblasts (red). The top 1,000 peaks in each condition where used to create the Naive T cell signature. Genes upregulated in the Naive T cell condition were given a value of +1, genes upregulated in the fibroblast condition were given a value of -1.

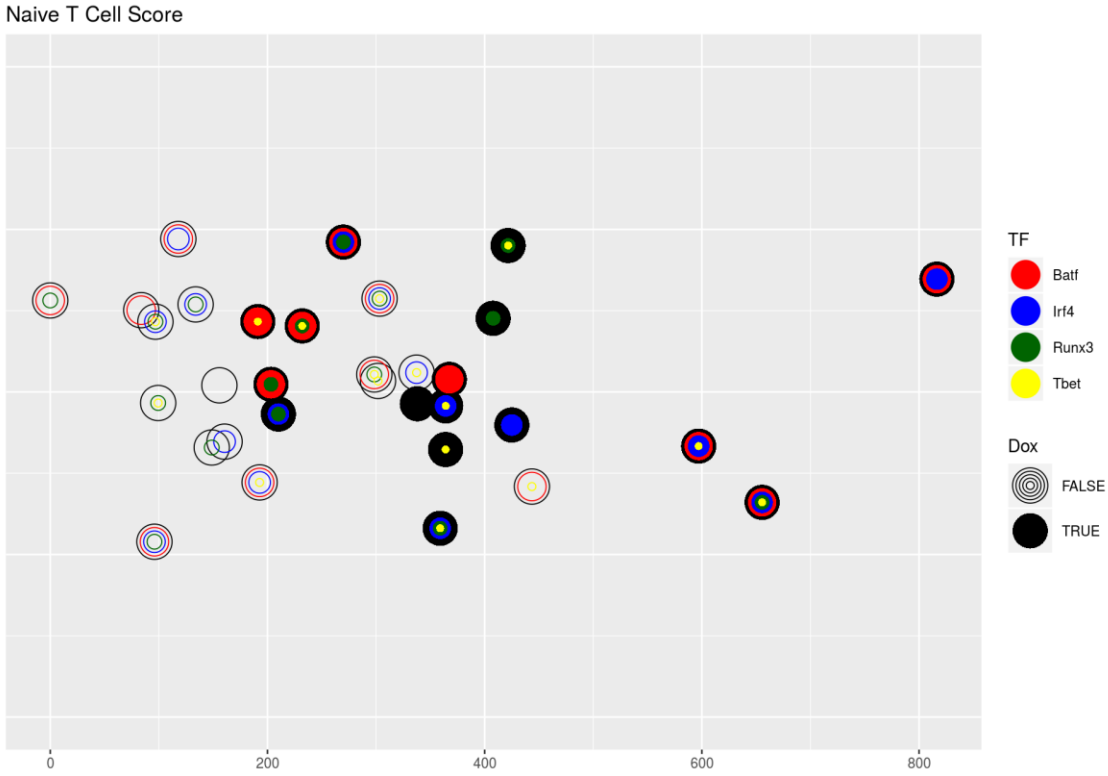
Figure 5F - ATAC-Seq Signature for Effector T Cells



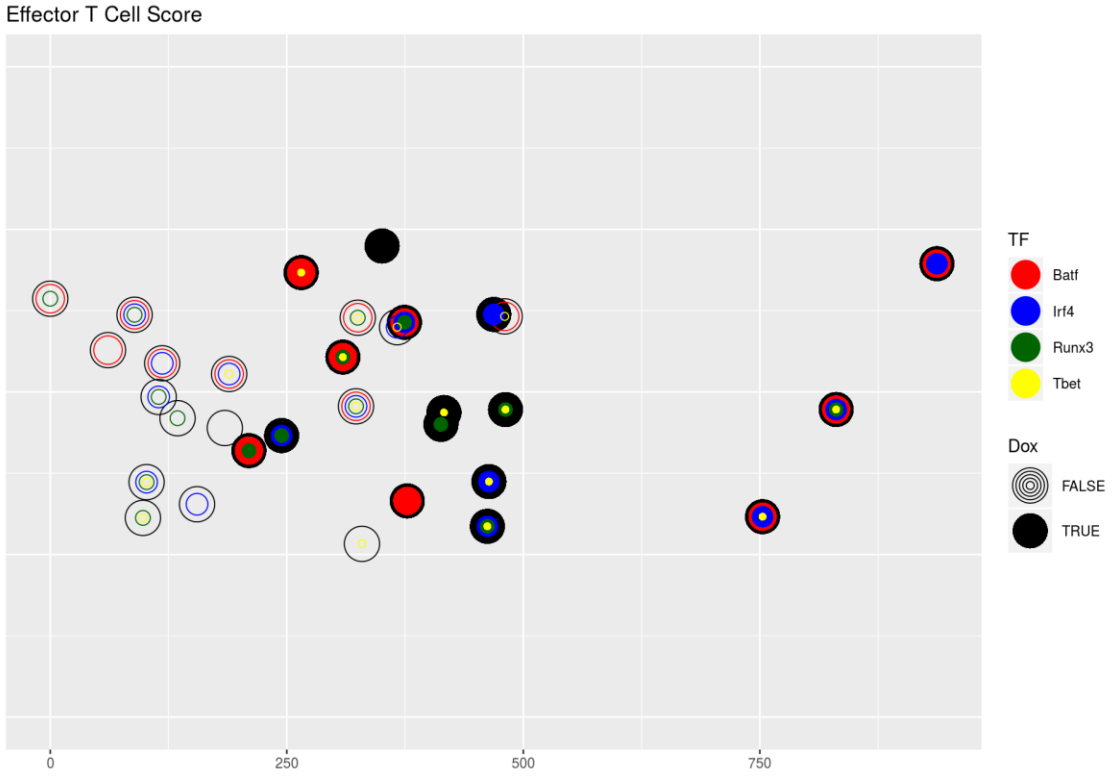
Legend: The volcano plot displays peaks that were differentially open in either in Effector T cell (green) or fibroblasts (red). The top 1,000 peaks in each condition where used to create the Effector T cell signature. Genes upregulated in the Effector T cell condition were given a value of +1, genes upregulated in the fibroblast condition were given a value of -1.

Figure 5G - T Cell Scores for the Fibroblast Experiments (ATAC)

A) Naive T Cell Score

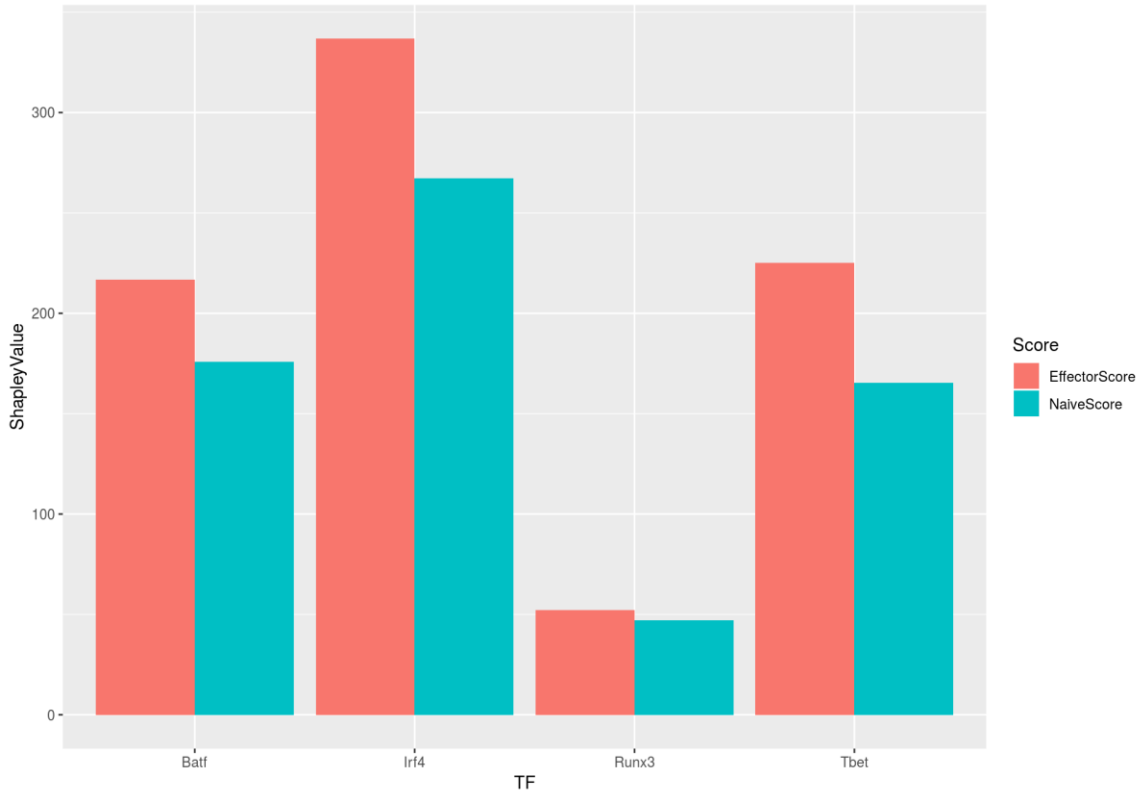


B) Effector T Cell Score



Legend: Each point represents a fibroblast sample. Empty circles represent samples without Doxocycline treatment, solid circles represent Doxocycline treatment. Colors indicate which of the four TFs have been added to the sample. The x-axis represents the T cell signature score. The y-axis is random "jitter" to separate points and make the plot easier to interpret.

Figure 5H - Shapley Value for Each TF for the ATAC-Seq T Cell Signatures



Legend: Each color bar represents the Shapley value, a measure of a TF's contribution to moving fibroblasts towards the given T cell signature.

4.8 References

Andris, Fabienne, Sébastien Denanglaire, Maelle Anciaux, Mélanie Hercor, Hind Hussein, and Oberdan Leo. 2017. "The Transcription Factor c-Maf Promotes the Differentiation of Follicular Helper T Cells." *Frontiers in Immunology* 8 (April): 480.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. "Trimmomatic: A Flexible Trimmer for Illumina Sequence Data." *Bioinformatics* 30 (15): 2114–20.

Buenrostro, Jason D., Paul G. Giresi, Lisa C. Zaba, Howard Y. Chang, and William J. Greenleaf. 2013. "Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-Binding Proteins and Nucleosome Position." *Nature Methods* 10 (12): 1213–18.

Glasmacher, Elke, Smita Agrawal, Abraham B. Chang, Theresa L. Murphy, Wenwen Zeng, Bryan Vander Lugt, Aly A. Khan, et al. 2012. "A Genomic Regulatory Element That Directs Assembly and Function of Immune-Specific AP-1-IRF Complexes." *Science* 338 (6109): 975–80.

Hahne F, Ivanek R. 2016. "Statistical Genomics: Methods and Protocols." In Mathé E, Davis S (eds.), chapter Visualizing Genomic Data Using Gviz and Bioconductor, 335–351. Springer New York, New York, NY. ISBN 978-1-4939-3578-9, doi: 10.1007/978-1-4939-3578-9_16, http://dx.doi.org/10.1007/978-1-4939-3578-9_16.

Kurachi, Makoto, R. Anthony Barnitz, Nir Yosef, Pamela M. Odorizzi, Michael A. DiIorio, Madeleine E. Lemieux, Kathleen Yates, et al. 2014. "The Transcription Factor BATF Operates as an Essential Differentiation Checkpoint in Early Effector CD8+ T Cells." *Nature Immunology* 15 (4): 373–83.

Langmead, Ben, and Steven L. Salzberg. 2012. "Fast Gapped-Read Alignment with Bowtie 2." *Nature Methods* 9 (4): 357–59.

Lareau, Caleb A., and Martin J. Aryee. 2018. "Hichipper: A Preprocessing Pipeline for Calling DNA Loops from HiChIP Data." *Nature Methods* 15 (February): 155.

Li, Bo, and Colin N. Dewey. 2011. "RSEM: Accurate Transcript Quantification from RNA-Seq Data with or without a Reference Genome." *BMC Bioinformatics* 12 (1): 323.

Liu, Ce Feng, Gabriel S. Brandt, Quyen Q. Hoang, Natalia Naumova, Vanja Lazarevic, Eun Sook Hwang, Job Dekker, Laurie H. Glimcher, Dagmar Ringe, and Gregory A. Petsko. 2016. "Crystal Structure of the DNA Binding Domain of the Transcription Factor T-Bet Suggests Simultaneous Recognition of Distant Genome Sites." *Proceedings of the National Academy of Sciences of the United States of America* 113 (43): E6572–81.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. "Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2." *Genome Biology* 15 (12): 550.

Lun, Aaron T. L., and Gordon K. Smyth. 2016. "Csaw: A Bioconductor Package for Differential Binding Analysis of ChIP-Seq Data Using Sliding Windows." *Nucleic Acids Research* 44 (5): e45.

McLean, Cory Y., Dave Bristor, Michael Hiller, Shoa L. Clarke, Bruce T. Schaar, Craig B. Lowe, Aaron M. Wenger, and Gill Bejerano. 2010. "GREAT Improves Functional

Interpretation of Cis-Regulatory Regions." *Nature Biotechnology* 28 (5): 495–501.
Mumbach, Maxwell R., Adam J. Rubin, Ryan A. Flynn, Chao Dai, Paul A. Khavari, William J. Greenleaf, and Howard Y. Chang. 2016. "HiChIP: Efficient and Sensitive Analysis of Protein-Directed Genome Architecture." *Nature Methods* 13 (11): 919–22.
Phanstiel, Douglas H., Alan P. Boyle, Nastaran Heidari, and Michael P. Snyder. 2015. "Mango: A Bias-Correcting ChIA-PET Analysis Pipeline." *Bioinformatics* 31 (19): 3092–98.
Servant, Nicolas, Nelle Varoquaux, Bryan R. Lajoie, Eric Viara, Chong-Jian Chen, Jean-Philippe Vert, Edith Heard, Job Dekker, and Emmanuel Barillot. 2015. "HiC-Pro: An Optimized and Flexible Pipeline for Hi-C Data Processing." *Genome Biology* 16 (December): 259.
Shan, Qiang, Zhouhao Zeng, Shaojun Xing, Fengyin Li, Stacey M. Hartwig, Jodi A. Gullicksrud, Samarchith P. Kurup, et al. 2017. "The Transcription Factor Runx3 Guards Cytotoxic CD8+ Effector T Cells against Deviation towards Follicular Helper T Cell Lineage." *Nature Immunology* 18 (8): 931–39.
Tanaka, Shigeru, Akira Suto, Taro Iwamoto, Daisuke Kashiwakuma, Shin-Ichiro Kagami, Kotaro Suzuki, Hiroaki Takatori, et al. 2014. "Sox5 and c-Maf Cooperatively Induce Th17 Cell Differentiation via RORγt Induction as Downstream Targets of Stat3." *The Journal of Experimental Medicine* 211 (9): 1857–74.
Zhang, Yong, Tao Liu, Clifford A. Meyer, Jérôme Eeckhoute, David S. Johnson, Bradley E. Bernstein, Chad Nusbaum, et al. 2008. "Model-Based Analysis of ChIP-Seq (MACS)." *Genome Biology* 9 (9): R137.