

COMPUTATIONAL PREDICTION OF SHRNA POTENCY AND ANALYSIS OF
CHROMATIN STATE TO DEFINE TUMOR-SPECIFIC T CELL DYSFUNCTION
AND REPROGRAMMABILITY

A Dissertation

Presented to the Faculty of the Weill Cornell Graduate School
of Medical Sciences
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

by

Lauren Elizabeth Fairchild

December 2018

COMPUTATIONAL PREDICTION OF SHRNA POTENCY AND ANALYSIS OF
CHROMATIN STATE TO DEFINE TUMOR-SPECIFIC T CELL DYSFUNCTION
AND REPROGRAMMABILITY

Lauren Fairchild, Ph.D.

Cornell University 2018

I present two computational studies of gene regulation: the first is a machine learning approach to predict potent shRNAs to knock down specific endogenous mRNAs; the second is an in-depth analysis of the relationship between chromatin accessibility and the functional state of tumor-specific CD8⁺ T cells.

Short-hairpin RNA, or shRNA, are synthetic RNA molecules that can be used to silence mRNA transcripts in a sequence-dependent manner and are used extensively in gain- and loss-of-function genetic studies. In order to predict which shRNA molecules will be potent despite changes in underlying shRNA technology, we developed a cascaded support vector machine, SplashRNA, trained on both types of shRNA backbone, miR-30 and miR-E. This strategy allows us to learn basic shRNA potency rules on our larger but older miR-30 dataset and followed by more precise miR-E-specific rules on our smaller miR-E training set. We demonstrate that SplashRNA outperforms all other shRNA prediction methods while limiting off-target effects through careful curation of mRNA transcript data and we have developed an open-source implementation of this algorithm available at splashrna.mskcc.org.

Tumor-specific T cells have been found in patients' tumors, but these tumors continue to progress, indicating that these T cells are not functional. A subset of patients in this situation have responded to checkpoint blockade therapy, which can rescue silenced T cells, but not all patients are responsive to this treatment and some only

respond for a brief period. Here, we investigate the chromatin accessibility landscape of normal and dysfunctional tumor-specific T cells in order to determine the regulatory changes that take place when T cells are in a dysfunction-inducing tumor environment. We computationally identify and pharmacologically validate NFAT and TCF family members as critical in this differentiation to dysfunction. We also identify cell surface markers that differentiate between reprogrammable and fixed cells, a strategy that may be used to determine if patients are candidates for checkpoint blockade therapy.

BIOGRAPHICAL SKETCH

Lauren Fairchild received her undergraduate degree in Biochemistry with a Computational focus from the University of Texas at Austin in 2012. During her time there she worked as an undergraduate research assistant developing RNA-seq and ChIP-seq methods under Professor Vishy Iyer and Dr. Patrick Killion. She joined the Tri-institutional training program in Computational Biology and Medicine in July of 2012 and joined Professor Christina Leslie's lab at the Sloan Kettering Institute in the fall of 2013. As a member of Prof. Leslie's lab, Lauren has developed computational methods for shRNA prediction and analysis of next-generation sequencing datasets.

ACKNOWLEDGEMENTS

Of course, I could not have performed any of this work alone. I would like to thank my research mentor, Christina Leslie, for her support and guidance during my graduate career. I also acknowledge and thank the members of the Leslie lab for their companionship, feedback, and discussions. I'd like to especially thank Rafi Pelosof for his advice and humor.

I thank my thesis committee: Olivier Elemento, Chris Mason, Michael Kharas for their thoughtful comments and questions.

I acknowledge and thank our collaborators Christof Fellmann, Mirimus, Andrea Schietinger, and Mary Philip. Without you this work would not have been possible.

I also acknowledge my undergraduate mentors, without whom I would not be where I am today. Thank you Patrick Killion for giving me a bright and hopeful start in research and being so willing to devote time and energy to your students.

I could not have accomplished this without the love and support of my family, specifically my parents, Mark and Debbie, and my stepmother Ann. I also would like to thank my extended family, especially Carol, Errol, and Linda, for making New Jersey my home away from home during my time here. Finally, thank you to my fiancé Sebastien for always being there through the highs and lows. We finally made it through – now on to our next chapter!

TABLE OF CONTENTS

BIOGRAPHICAL SKETCH	III
ACKNOWLEDGEMENTS	IV
LIST OF TABLES	IX
CHAPTER 1 INTRODUCTION	1
1.1 INTRODUCTION TO GENE REGULATION	1
1.2 INTRODUCTION TO SHRNA TECHNOLOGY	2
1.2.1 <i>History of RNA interference</i>	2
1.2.2 <i>A Sensor assay to test potency of large shRNA libraries</i>	3
1.2.3 <i>shRNA backbones</i>	4
1.2.4 <i>Existing methods of shRNA prediction</i>	5
1.3 T CELL INTRODUCTION	6
1.3.1 <i>Introduction to CD8+ T cell function</i>	6
1.3.2 <i>Tumor-specific T cells and checkpoint blockade therapy</i>	7
1.3.3 <i>A mouse model of tumor-specific dysfunctional CD8+ T cells</i>	9
1.3.4 <i>Chromatin profiling to determine functional state of CD8+ T cells</i>	11
CHAPTER 2 PREDICTION OF POTENT SHRNA WITH A SEQUENTIAL CLASSIFICATION ALGORITHM	14
2.1 INTRODUCTION.....	14
2.2 A SEQUENTIAL CLASSIFIER TO EXPLOIT DIVERSE RNAI DATASETS	16
2.3 TRAINING THE COMPONENT SPLASHRNA CLASSIFIERS.....	21
2.4 SPLASHRNA OUTPERFORMS EXISTING SHRNA PREDICTION METHODS	24
2.5 TARGETING THE RELEVANT TRANSCRIPT SPACE	30
2.6 <i>IN VIVO</i> VALIDATION OF <i>DE NOVO</i> PREDICTED SHRNAs	34

2.7	MINIMIZATION OF OFF-TARGET EFFECTS	39
2.8	DISCUSSION.....	41
CHAPTER 3 CHROMATIN STATES DEFINE TUMOR-SPECIFIC T CELL		
DYSFUNCTION AND REPROGRAMMING		
3.1	INTRODUCTION.....	42
3.2	CD8 T CELL CHROMATIN CHANGES DURING INFECTION	43
3.3	CHROMATIN STATE DYNAMICS OF TST DYSFUNCTION	47
3.4	CHROMATIN STATES CORRELATE WITH REPROGRAMMABILITY	51
3.5	SURFACE PROTEINS ASSOCIATED WITH CHROMATIN STATES	56
3.6	MEMORY T CELLS ENTER FIXED DYSFUNCTIONAL STATE IN TUMOR.....	60
3.7	CHROMATIN ACCESSIBILITY IN HUMAN TILS	63
3.8	DISCUSSION.....	66
3.9	METHODS.....	67
CHAPTER 4 MODELING OF TRANSCRIPTION FACTOR ACTIVITIES TO		
PREDICT CHANGES IN GENE EXPRESSION		
4.1	INTRODUCTION.....	74
4.2	T CELL MATURATION AND SELF-TOLERANCE MECHANISMS	74
3.3	COMPUTATIONAL MODELING OF TWO DYSFUNCTIONAL T CELL STATES.....	75
CHAPTER 5 CONCLUSIONS AND FUTURE DIRECTIONS		
5.1	DISCUSSION.....	84
5.2	shRNA POTENCY PREDICTION.....	84
5.3	T CELL EPIGENETIC MODELING	86
5.4	CONCLUSION.....	89
BIBLIOGRAPHY.....		
		90

LIST OF FIGURES

Figure 1.1: shRNA sensor construct.	3
Figure 1.2: Polypeptides presented on cell surface by MHC class I pathway.	8
Figure 1.3: Tamoxifen-inducible liver cancer model.	10
Figure 1.4: Experimental design.	11
Figure 1.5: Schematic of ATAC-seq procedure.	12
Figure 1.6: Representative ATAC-seq library insert-size distribution.	13
Figure 2.1: Computational modeling of advancements in shRNA technology.	15
Figure 2.2: Nucleotide representation of positive shRNAs from the indicated datasets.	18
Figure 2.3: Kernel selection.	19
Figure 2.4: Incorporation of M1 dataset to generate a miR-30 classifier.	20
Figure 2.5: Calibration of the sequential SVM classifier SplashRNA.	23
Figure 2.6: Performance of various shRNA prediction algorithms.	27
Figure 2.7: SplashRNA performance on <i>in vivo</i> screens.	28
Figure 2.8: Transcript selection.	32
Figure 2.9: Western blot validation of <i>de novo</i> SplashRNA predictions.	35
Figure 2.10: SplashRNA comparison to CRISPR-Cas9.	37
Figure 3.1: Flow cytometric analysis of characteristic markers after <i>LmTAG</i> stimulation.	43
Figure 3.2: Chromatin accessibility in normal CD8 T cells.	44
Figure 3.3: Activation of early response genes.	45
Figure 3.4: Linked k-means clustered RNA-seq and ATAC-seq heatmaps.	46

Figure 3.5: Flow cytometric analysis of inhibitory markers in TST.....	47
Figure 3.6: Overview of chromatin accessibility in TST.....	49
Figure 3.7: ATAC-seq signal in <i>Pdcd1</i> and <i>Ifng</i> loci in normal and TST cells.....	50
Figure 3.8: Transcription factor accessibility in naïve to L5 transition.....	51
Figure 3.9: Tumor-associated T cell dysfunction occurs in two states.....	52
Figure 3.10: Pharmacological targeting of NFAT and Wnt/ β -catenin signaling prevents TST differentiation to the fixed dysfunctional state <i>in vivo</i>	54
Figure 3.11: Cell surface markers for dysfunctional state transition.....	58
Figure 3.12: ATAC-seq signal profile across the <i>Cd38</i> locus.	59
Figure 3.13: B16-OVA model	59
Figure 3.14: Memory CD8 T cells in established tumors.....	61
Figure 3.15: Dysfunctional memory cells.....	62
Figure 3.16: Normal human T cells and tumor-infiltrating lymphocytes.....	63
Figure 3.17: Comparison of mouse and human chromatin accessibility	65
Figure 3.18: Characterization of normal human T cells and TILs.....	66
Figure 4.1: PCA of CD8 ⁺ T cell types.	77
Figure 4.2: Model performance scatterplots.	79
Figure 4.3: Heatmaps of learned transcription factor weights scaled by cisBP-derived binding terms.	82

LIST OF TABLES

Table 2.1: shRNA potency datasets used for training and performance assessment.....	17
Table 2.2: Polyadenylation sites in Pten gene affect observed shRNA knockdown.	31
Table 3.1: GO terms enriched in regions closing in L7 to L14 transition	53
Table 3.2: Number of peaks per cell type	69
Table 4.1: Samples used in modeling	76
Table 4.2: Model performance across six cell type transitions.....	80
Table 4.3: Influential transcription factors identified by model	81

CHAPTER 1

INTRODUCTION

1.1 Introduction to gene regulation

Although humans¹ and mice² have about the same number of protein-coding genes as the relatively simple nematode *Caenorhabditis elegans* (~ 20,000 genes³), and about four times as many genes as budding yeast *Saccharomyces cerevisiae* (~6,000 genes⁴), the regulation of those genes are orders of magnitude more complex⁵. Completing the sequence of the first human genome in the early 2000s ushered in a new age of biological research. It was discovered that although some diseases are caused by single mutations in coding regions that are relatively easy to identify, many more cannot be explained by a single mutation and are caused by disruptions in complex gene regulatory networks or inappropriate responses to environmental cues. As genomic and transcriptomic regulation are highly relevant in the clinic and for our understanding of biology, questions of gene regulation have been approached from many angles including chromatin organization, epigenetic signatures, alternative mRNA splicing, alternative 5' and 3' untranslated regions, intron retention, and more.

In this work, I describe two studies that attempt to address different questions in gene regulation. The first is a machine learning approach, SplashRNA, designed to predict potent short-hairpin RNAs to knock down endogenous mRNA transcripts in a sequence specific manner. The second is an epigenetic and transcriptomic profiling of tumor-specific CD8⁺ T cells (TSTs) to determine the mechanisms of dysregulation in TSTs relative to normal cytotoxic T-cells.

1.2 Introduction to shRNA technology

1.2.1 History of RNA interference

Focused and large-scale functional genomics approaches performed in mammalian cells and model organisms have the potential to uncover gene interaction networks for the better understanding of homeostasis and disease, and the development of novel therapeutics. For over a decade, RNA interference (RNAi) has been the technology of choice for both positive and negative selection screens in higher eukaryotes and has provided unparalleled insight into gene function.

RNAi provides a programmable mechanism for reversible suppression of gene expression⁶. Through a highly conserved pathway, the RNAi machinery processes double-stranded RNAs into small RNAs that guide the repression of complementary genes [reviewed in ⁷]. Experimental RNAi acts by providing exogenous sources of double-stranded RNA that mimic endogenous triggers to enable rapid gene knockdown. Importantly, the single component nature of RNAi tools makes them extremely amenable to large-scale applications and delivery in nearly any model system. RNAi triggers can be designed to either inhibit specific messenger RNAs (mRNAs) or suppress all splice isoforms through targeting of the common transcript.

However, perfect sequence complementarity is neither necessary nor sufficient for strong RNAi knockdown⁸. shRNA may bind to off-target sites with few mismatches and may not bind sites even if there is a perfect sequence match. There are also sequence restrictions of the endogenous miRNA processing machinery such that some sequences are more efficiently processed than others. Because not all the sequence restraints for shRNA and other RNAi technologies have been elucidated, there is still a relatively low rate of success of shRNAs. Usually only about 50% of predicted shRNAs

are potent using existing prediction tools, leading to a significant amount of lost time and resources⁹.

1.2.2 A Sensor assay to test potency of large shRNA libraries

To better understand the sequence requirements for potent RNAi and identify efficient microRNA-embedded shRNAs for any gene, we have developed a functional high-throughput “Sensor” assay that allows biological assessment of ten thousands of shRNAs in parallel¹⁰. This assay relies on a “Sensor” construct where the expression of the shRNA is driven by a Tet-inducible promoter. Also a part of this construct is the Venus gene, which codes for a yellow-fluorescent protein (YFP), under a constitutively active promoter. The cognate sequence of the shRNA is included in the 3' UTR of the Venus sequence and acts as a target for the shRNA. If the shRNA is potent and expressed, it will bind to this target sequence and the Venus RNA will be targeted for degradation, decreasing the amount of fluorescence detected. Thus, a pool of cells containing a library of shRNAs can be sorted using fluorescence activated cell sorting

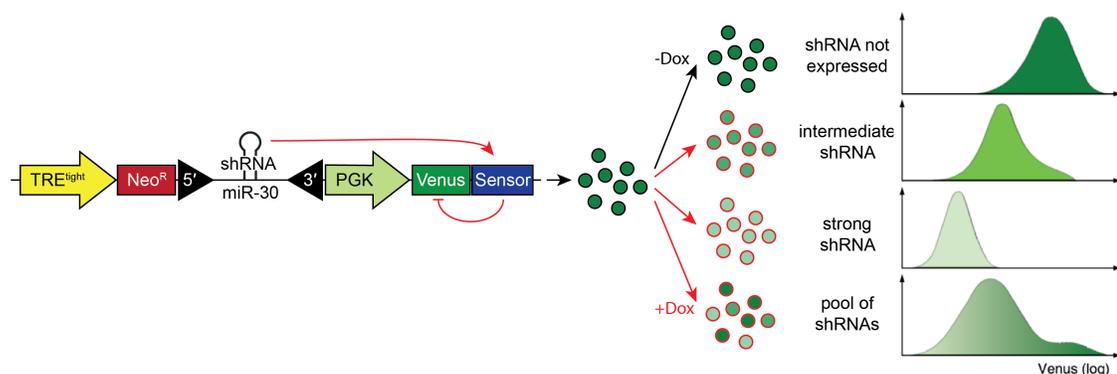


Figure 1.1: shRNA sensor construct.

The target Sensor sequence harbors the reverse-complement of the Tet-inducible shRNA and is located in the 3'UTR of the constitutively active Venus reporter. Venus expression then directly reports shRNA potency. Figure is adapted from Figure 1a of Fellmann et al, Mol Cell, 2011.

(FACS) by their repression of Venus (**Figure 1.1**, adapted from ¹⁰). Multiple rounds of sorting on- and off-doxycycline leads to a purified population of cells containing constructs with potent shRNA. The constructs of these cells are then sequenced to determine the shRNA identities.

We initially applied this Sensor assay to evaluate 20,000 shRNAs at each step of microRNA biogenesis. To design these sequences, we completely tiled nine genes. This allowed us to uncover in an unbiased manner both known and novel processing step-specific features of potent RNAi. We also used this reporter assay to generate focused and genome-wide shRNA libraries^{11,12}, biologically testing well over 300,000 microRNA-embedded shRNAs overall. While the Sensor assay robustly identified the most potent sequences, generating large-scale libraries using this approach was time consuming and impractical as it became a balance between reducing the number of genes in a library versus reducing the number of candidate shRNAs per gene to be tested. Yet, the Sensor assay generated invaluable datasets that when integrated correctly can serve as ideal training and validation sets for computational shRNA prediction algorithms.

1.2.3 shRNA backbones

The backbone of an shRNA is the construct into which the targeting shRNA sequence is placed. As an shRNA is processed by the cell's miRNA production pathway, the backbone must be similar enough to endogenous sequences to be recognized. For this reason, a commonly used shRNA backbone was based upon the human MIR30A sequence.

To increase the potency of all shRNAs, especially when expressed from a single-copy genomic integration, we have established an optimized microRNA backbone (miR-E)

that boosts processing efficiency 10-30 fold when compared to the standard miR-30 backbone, leading to much stronger target knockdown in most cases¹³. The miR-E scaffold restores a conserved 5'-DCNNC-3' motif in the 3'-flank of the backbone that is recognized by components of the microprocessor complex for efficient biogenesis, similar to most endogenous microRNAs^{13,14}. This conserved sequence had been replaced with an EcoRI restriction site in the miR-30 backbone, leading to decreased shRNA processing efficiency. Although this change increased the overall efficiency of shRNAs, it increased the difficulty of potent shRNA prediction as the miR-E reagents behaved slightly differently than their miR-30 predecessors.

1.2.4 Existing methods of shRNA prediction

Algorithmic developments and prediction performance are closely tied to the quality, size and implementation of the (siRNA or shRNA) datasets used for training. While initial rules of RNAi potency contained many non-sequence features derived from relatively low-throughput studies¹⁵⁻¹⁷, later rules inferred from larger screens found that sequence based features are more predictive^{18,19} and indirectly capture the other characteristics²⁰. As one of the larger early studies, the BIOPREDSi algorithm, a non-linear neural network, was trained on over 2,000 functionally tested siRNAs targeting 34 genes and set a new performance standard. Using the same dataset, the DSIR algorithm improved performance through the use of an L1 regularized linear model with a combination of a degree 3 spectrum-kernel and sequence and position features^{19,21}. However, it soon became apparent that the rules governing shRNA efficiency differ from the ones dictating siRNA potency^{10,22}, explaining why siRNA based algorithms performed relatively poorly in shRNA prediction tasks. This difference is likely due to

the additional processing requirements imposed on stem-loop shRNAs by the Dicer complex, and on microRNA-embedded shRNAs by the Drosha and Dicer complexes.

Thus, we and others have previously used our large-scale microRNA-based shRNA potency datasets to generate miR-30 specific prediction algorithms^{12,23}. However, with a shift towards the use of the much more efficiently processed miR-E backbone¹³, these algorithms are no longer accurately trained for the task at hand as key sequence requirements have changed. In chapter 2, I describe the development of SplashRNA, a cascaded support vector machine classifier for predicting potent miR-E based shRNA that outperforms all existing prediction methods.

1.3 T cell introduction

1.3.1 Introduction to CD8+ T cell function

Although there are many cell types involved in the immune system and immune response, here we will focus on CD8+ or cytotoxic T cells. These cells are able to kill invading pathogens when activated by a major histocompatibility complex (MHC) protein presenting their specific antigen (reviewed in ²⁴). During an infection with a bacteria or virus, antigen-naive T cells will be activated by an antigen-presenting cell presenting their specific antigen, often a component of the invading pathogen. Activated T cells differentiate into effector cells, undergo clonal expansion, and search for their specific antigen. When infected target cells presenting this antigen are located, the effector T cells bind to the target cell's surface and release cytotoxins, triggering apoptosis.

1.3.2 Tumor-specific T cells and checkpoint blockade therapy

Cytotoxic T cells are also able to play a role protecting the body by killing cancerous cells. During cancer's progression, even during pre-malignant stages, both cancer-inducing "driver" mutations and benign "passenger" mutations frequently occur²⁵. If these mutations occur within the coding sequence of a gene, the protein product may have an altered polypeptide sequence as a result. As part of the major histocompatibility complex (MHC) class I antigen presentation pathway, cytosolic proteins are degraded and portions of these proteins, called polypeptides, are presented on the cell surface as part of a MHC-peptide complex. Polypeptides containing a cancer-induced alteration may be presented through this pathway and as these molecules are not endogenous, T cells with antigens specific to these polypeptides will not have been eliminated by central or peripheral self-tolerance mechanisms²⁶. Therefore, it is possible for a T cell to recognize and bind the altered polypeptide and target the presenting cell for destruction (**Figure 1.2**, adapted from ²⁶).

Cancer cells may avoid this type of immune targeting by both creating an immunosuppressive microenvironment and direct suppression of T cell activity²⁷. This suppression is achieved by aberrantly expressing ligands for immune checkpoint molecules, the most well-studied of which are CTLA-4 (cytotoxic T-lymphocyte antigen 4) and PD-1 (programmed death 1). These molecules are expressed in normal cells to prevent autoimmunity. Several biologic therapies have been developed to target and inhibit these and other immune checkpoint molecules, a strategy termed checkpoint blockade (reviewed in ²⁸). Some of these drugs are monoclonal antibodies that bind to CTLA-4 (ipilimumab, the first approved checkpoint blockade drug) or PD-1 (Nivolumab and Pembrolizumab) and interrupt the interaction between these receptors and their ligands. Blocking these interactions leads to increased T cell activity and has

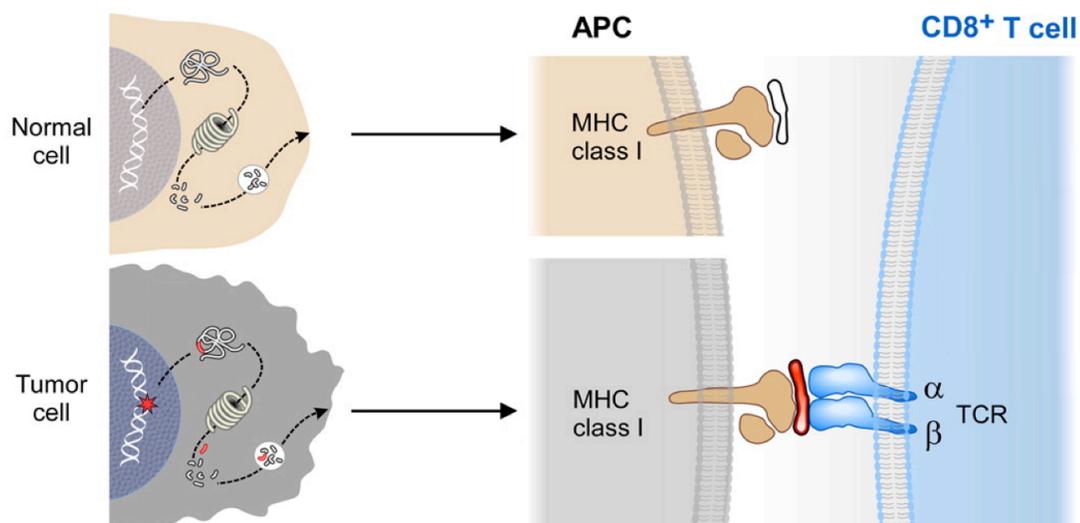


Figure 1.2: Polypeptides presented on cell surface by MHC class I pathway. Mutations in tumor cells (red, bottom) may lead to altered polypeptides being presented. As these are non-native polypeptides, CD8+ T cells may recognize and bind these presented tumor neoantigens. (Figure adapted from Fritsch, Hacohen & Wu, *OncoImmunology* 2014)

been shown to be effective in patients with advanced and metastatic cancers including lung cancer and melanoma.

These checkpoint blockade molecules have not been a panacea, however. As they activate immune cells globally, many patients treated with these drugs develop autoimmune reactions such as colitis or hepatitis²⁸. More troubling however, is that not all patients with cancers expressing PD-L1 and CTLA-4 ligands respond to these treatments. Developing methods to determine which patients will respond to immune checkpoint blockade therapy and how to induce a therapeutic response in the remaining population are research areas of critical importance in oncology research and will be addressed in this work.

Even when immune checkpoint mechanisms have been blocked, T cells do not always activate and attack their target cancer cells. Additionally, many patients experience resistance to this therapy after a relatively short period of time.²⁸ This persistent T cell dysfunction is similar to the T cell exhaustion observed in chronic viral infection. T cells in both the tumor environment and in chronic viral infection have

decreased effector function and express inhibitory molecules such as PD-1, LAG-3, TIM-3, and CTLA-4 and have an altered transcriptional state relative to normal effector T cells in response to persistence of their specific antigen²⁹. One hypothesis is that this altered phenotypic and expression state is accompanied and perhaps driven by a correspondingly altered epigenetic state. In chapter 3, we investigate how the epigenetic state in tumor-specific T cells is altered even in pre-malignant lesions, before the tumor microenvironment is established and how this correlates with reprogrammability of tumor-specific T cells.

1.3.3 A mouse model of tumor-specific dysfunctional CD8⁺ T cells

In order to study the progression of dysfunction in tumor-specific CD8⁺ T cells (TST), we used a tamoxifen-inducible, autochthonous liver cancer model (AST-Cre-ER^{T2}) in which the SV40 large T antigen (TAG) acts as both the tumor-initiating factor and the tumor-specific antigen³⁰. The AST cassette consists of three components: the hepatocyte-specific albumin promoter, a loxP flanked stop-cassette, and the SV40 large T antigen (**Figure 1.3**). Under normal conditions, mice with this construct will not express the T antigen due to the presence of the stop cassette. When Cre is added to the system or expressed endogenously, the stop-cassette between the loxP sites is deleted and the T antigen is expressed. Mice with the AST construct under a liver-specific promoter are crossed with mice containing a tamoxifen-inducible Cre gene, allowing the deletion of the stop-cassette and expression of the SV40 large T antigen to be controlled through administration of tamoxifen in the offspring. This experimental scheme allows for studies in early tumor development and differentiation along with studies of TAG-specific T cell (TCR_{TAG}) efficacy in early tumorigenesis, as the exact day of tumor initiation is known.

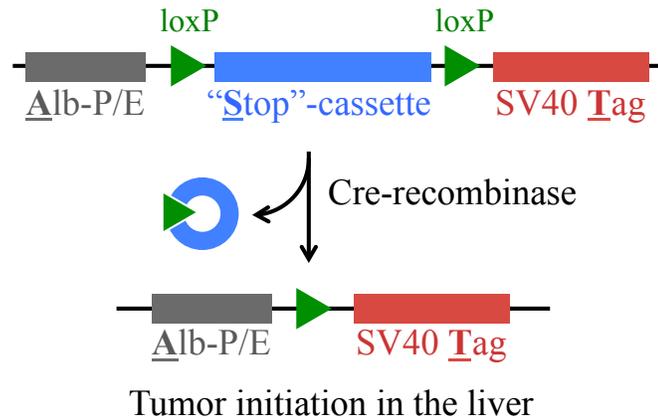


Figure 1.3: Tamoxifen-inducible liver cancer model.

Tam-induced Cre-mediated excision of the flox-stop cassette leads to SV40 large T antigen expression.

To accompany this study of TSTs, we also transferred congenically marked naive (N; CD44^{lo}CD62L^{hi}) TCR_{TAG} cells into wild-type C57BL/6 mice. These wild-type mice were then immunized one day later with a recombinant *Listeria monocytogenes* strain expressing this same TAG epitope (*LmTAG*)^{7,14}. This led to a normal immune response in which the naive TCR_{TAG} cells recognized and bound to the TAG epitope, activated and differentiated to effector cells, and then further differentiated to memory cells once the *LmTAG* infection was cleared. This is a controlled system for studying the progression of T cells in a normal environment.

As shown in **Figure 1.4**, T cells were extracted from both the normal model and the tumor model at various time points. We then performed both RNA-seq and ATAC-seq on these cells to profile them transcriptionally and epigenetically.

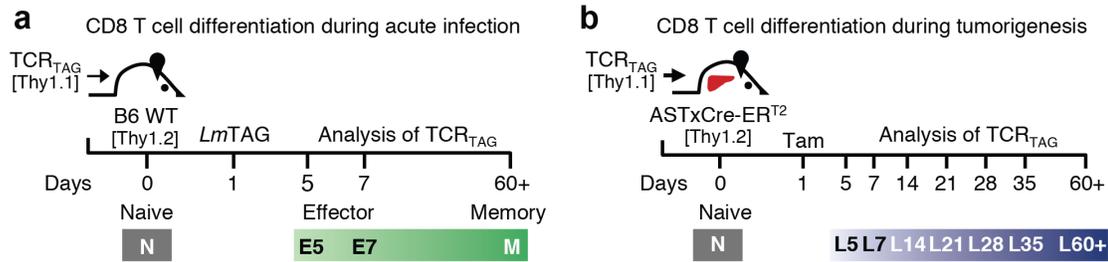


Figure 1.4: Experimental design.

Experimental design in both the acute infection model (a) and during tumorigenesis (b). Days indicate time points at which TCR_{TAG} cells were isolated and analyzed.

1.3.4 Chromatin profiling to determine functional state of CD8+ T cells

As every cell contains the complete genetic code of the organism, spanning about 3 billion bases in humans and mouse^{1,2}, DNA is, by necessity, tightly compacted. The organization of DNA folding is tightly regulated and this regulation partially determines the genetic programs activated. Regions of DNA that are folded less tightly are accessible to DNA binding proteins including RNA polymerase and transcription factors which regulate gene transcription both proximally and distally³¹.

ATAC-seq (Assay for Transposase Accessible Chromatin followed by sequencing)³² is a genome-wide high-throughput sequencing method for detecting regions of open chromatin, similar to DNase-seq. However, ATAC-seq can be performed with many fewer cells than DNase-seq allowing for profiling of less-common cell populations. The critical step of ATAC-seq is the addition of a transposase (Tn5) loaded with sequencing adapters to the cells of interest. In less tightly compacted regions of DNA where the Tn5 transposase is able to bind, it will cleave the DNA and insert its adapters into the genome. This combines the genome fragmentation and tagging steps of a standard sequencing library preparation protocol into a single

“tagmentation” step. These fragments of DNA are then purified, sequenced, and mapped back to the genome to determine which regions were accessible at the time when the transposase was added (**Figure 1.5**, adapted from Figure 1a of ³²).

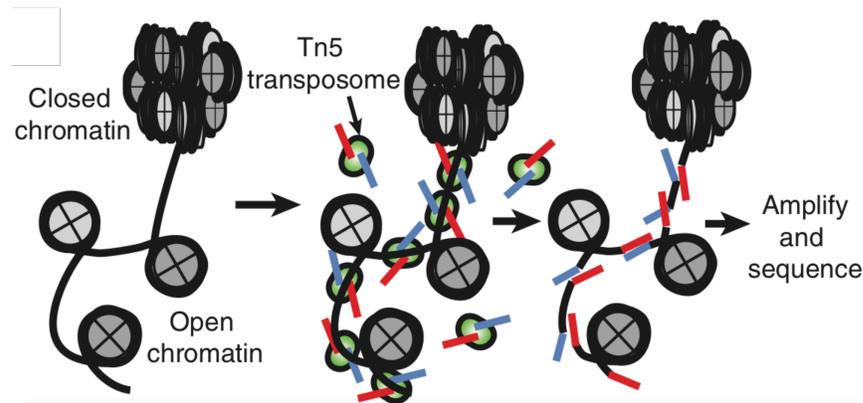


Figure 1.5: Schematic of ATAC-seq procedure.

Tn5 transposase loaded with sequencing adapters (red, blue) is added to chromatin. In regions where the transposase is able to bind, the adapters are inserted and DNA is cleaved. The resulting fragments are purified, amplified, and sequenced. (Figure adapted from Buenrostro et al, Nat Methods 2013)

Paired-end ATAC-seq sequencing libraries have a distinctive insert size distribution due to two factors: the 10.5bp helical pitch of B-DNA, the most common helical structure of DNA, and the spanning of DNA across nucleosomes. DNA that is wrapped around a nucleosome protected from cleavage by the transposase, leading to an enrichment of reads with lengths that are multiples of about 146-147 bases, the length of DNA protected by the nucleosome (**Figure 1.6**)³³. Checking for the presence of these two markers is one quality control step in the ATAC-seq library preparation.

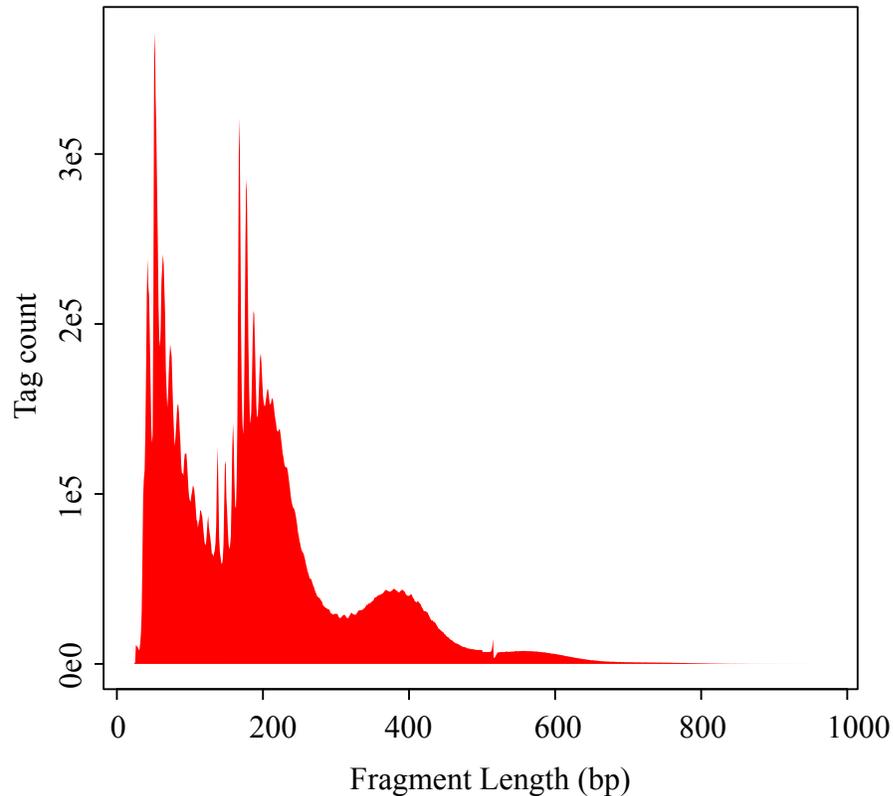


Figure 1.6: Representative ATAC-seq library insert-size distribution. Tag count enrichments at multiples of 178bp represent nucleosome-protected tags. Smaller peaks occurring every 10.5bp indicate the 10.5bp helical pitch of DNA.

In addition to and perhaps more important than knowledge of nucleosome positioning, regions of open chromatin indicate potential sites of transcription factor binding and regulation. It is not feasible to experimentally test for all DNA binding protein occupancies, but we can infer the binding of these proteins using sequence motifs. By comparing the known binding motifs of transcription factors to the sequence under the summit of reproducible peaks, the probability that a transcription factor is bound to a peak region can be estimated. By comparing genome-wide changes in transcription factor binding and occupancy, global changes in transcription factor activity can be inferred. This analysis is performed in chapter 3 to compare transcription factor activities in normal and dysfunctional CD8⁺ T cells.

CHAPTER 2

PREDICTION OF POTENT shRNA WITH A SEQUENTIAL CLASSIFICATION ALGORITHM

Portions of this chapter first appeared in Fairchild* et. al.⁹ and were written in collaboration with Raphael Pelossof, Christof Fellmann, and Christina Leslie.

2.1 Introduction

Experimental RNA interference (RNAi) acts by providing exogenous sources of double-stranded RNA that mimic endogenous triggers and enable reversible, transcript-specific gene knockdown⁷. Whereas short interfering RNAs (siRNAs) allow for rapid gene knockdown, they are not suitable for many long-term and in vivo studies due to their transient nature. Stem-loop shRNAs can be used as a continuous source of RNAi triggers when expressed from suitable vectors, but suffer from various technical limitations including inaccurate processing³⁴ and off-target effects through saturation of the endogenous microRNA machinery³⁵⁻³⁷. State-of-the-art microRNA-based shRNA vectors can overcome these limitations by providing a natural substrate of the RNAi pathway that is accurately and efficiently processed^{13,38-40} resulting in minimal or no off-target effects when expressed from a single genomic integration (single-copy)¹⁰.

Still, our limited understanding of RNAi processing requirements and the lack of robust algorithms for the design of microRNA-based shRNAs with high potency and low off-target activity has hampered the utility of RNAi tools. To understand the sequence requirements of potent RNAi and identify efficient microRNA-based shRNAs for any gene, we previously developed a functional high-throughput “Sensor” assay that enables biological assessment of tens of thousands of shRNAs in parallel (**Figure 2.1a**)¹⁰. We used this assay to generate focused and genome-wide shRNA libraries^{11,12}.

Furthermore, to increase the potency of all shRNAs, especially when expressed at single-copy, we established miR-E, an optimized microRNA backbone that boosts processing efficiency^{13,14} and leads to stronger target knockdown when compared to standard miR-30 designs¹³.

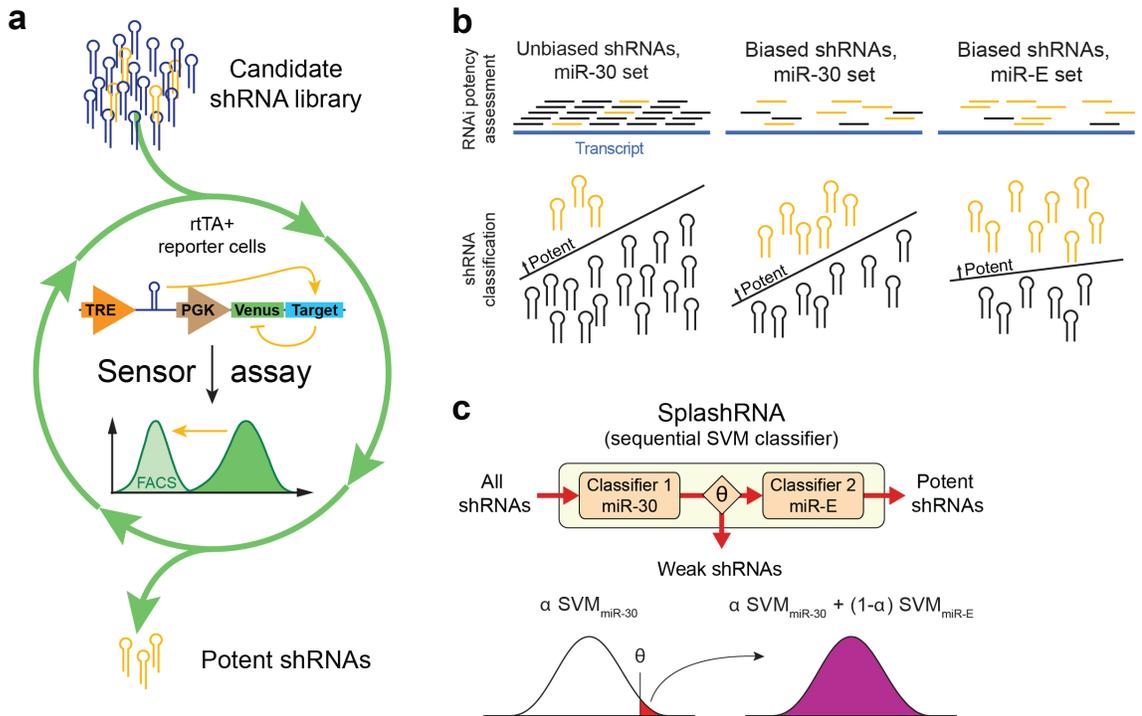


Figure 2.1: Computational modeling of advancements in shRNA technology. (a) Schematic of our previously published Sensor assay that enables large-scale functional assessment of shRNA potency. (b) Schematic of diverse biological shRNA potency datasets and their feature (top) and class label (bottom) distribution biases. Unbiased large-scale sets include a comprehensive representation of negatives but contain few potent shRNAs (left panel, potent=yellow). Sets selected using prediction tools show a higher rate of positives, at the cost of changing the feature distribution of the negatives (middle panel). Use of the optimized miR-E backbone changes the requirements for potent RNAi, altering the target prediction rule (right panel). (c) Concept and equation of SplashRNA. We model the advancement in shRNA technology as a sequential support vector machine (SVM) classifier. The first classifier is trained on miR-30 data and removes non-functional sequences and the second classifier is trained on miR-E data to increase prediction performance of the remaining shRNAs in the miR-E setting. The final output is a weighted combination of the scores from both classifiers.

The performance of shRNA prediction tools depends on the quality, size and design of the datasets used for training. While initial rules of RNAi potency contained many non-sequence elements¹⁵⁻¹⁷, later rules inferred from larger screens found that sequence based features are more predictive^{18,19} and capture the other characteristics²⁰.

BIOPRED*si*, a neural network approach, was trained on over 2,000 functionally tested siRNAs and set a new performance standard¹⁸. Using the same training dataset, DSIR improved prediction through the use of an L1 regularized linear model with a combination of position-specific nucleotide features and mono-, di-, and tri-nucleotide counts^{19,21}. However, the rules governing siRNA potency differ from the ones dictating shRNA potency due to the additional biogenesis steps^{10,22}, and siRNA-based algorithms perform relatively poorly in shRNA prediction tasks. Hence, we and others have previously used our large-scale shRNA datasets to generate miR-30 specific prediction algorithms^{12,23}. Still, with a shift towards the more efficiently processed miR-E backbone, these algorithms are no longer designed for the task at hand as key sequence requirements have changed.

2.2 A sequential classifier to exploit diverse RNAi datasets

To build an accurate miR-E shRNA predictor, we developed SplashRNA, a sequential learning algorithm combining two support vector machine (SVM) classifiers trained on judiciously integrated data sets (**Table 2.1**). SplashRNA models the sequential advances in shRNA technology to enable efficient learning on unbiased and biased data (**Figure 2.1b,c**). To train the algorithm, we generated a large-scale miR-30 data set (referred to as M1); and a miR-E data set (referred to as miR-E) using our RNAi Sensor and reporter assays, respectively^{10,13}. We also incorporated the previously published TILE¹⁰ and UltramiR¹² sets. The UltramiR dataset uses shRNAs constructed using the UltramiR

backbone. This backbone has been shown to be functionally indistinguishable from the miR-E backbone¹².

Table 2.1: shRNA potency datasets used for training and performance assessment
The total count of shRNAs in each library is indicated (N) along with the number of positive (N-pos) and negative (N-neg) examples

	Backbone	Screen type	N	N-pos	N-neg	Use
TILE	miR-30	Sensor assay, pooled	18720	5736	12685	Training, validation
M1	miR-30	Sensor assay, pooled	20324	9602	10722	Training, validation
mRas + hRAS	miR-30	Sensor assay, pooled	9804	1139	8665	Validation
shERWOOD 250k	miR-30	Sensor assay, pooled	227673	53234	174439	Validation
miR-E	miR-E	Reporter assay, one-by-one	397	170	227	Training, validation
UltramiR	UltramiR*	Cell viability, pooled	780	378	402	Training, validation
Essential genes, Top50 hits	Mini miR-30 with DCNNC motif*	Cell viability, pooled	1002			Validation
Sensitivity genes, Top20 hits	Mini miR-30 with DCNNC motif*	Toxin resistance and sensitivity, pooled	500			Validation

* These miRNA-based shRNA backbones are functionally equivalent to miR-E.

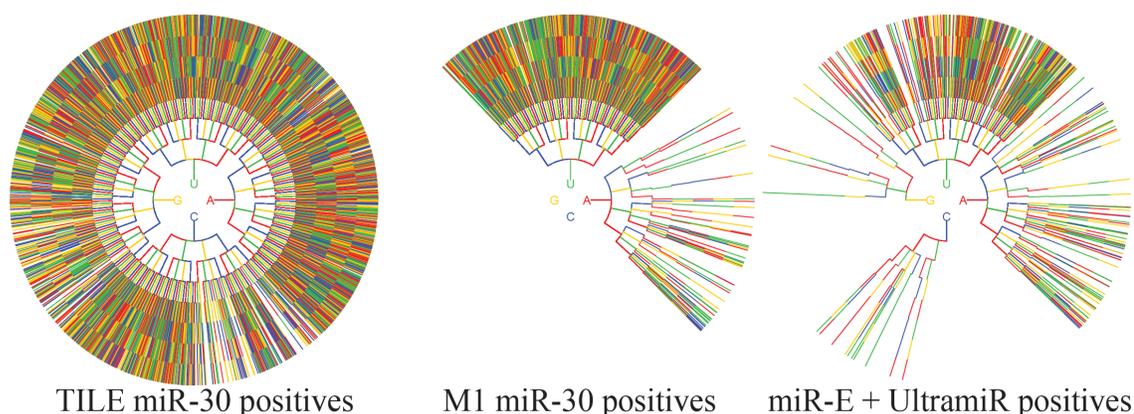


Figure 2.2: Nucleotide representation of positive (potent) shRNAs from the indicated datasets.

Shown are the nucleotides one to eight of the guide strand (starting in the center), including the entire seed region. Unbiased TILE (miR-30) set, showing a diverse nucleotide composition (left panel). Preselected M1 (miR-30, selected by DSIR + sequence rules derived from sensor assay) set, showing a biased nucleotide representation (middle panel). Preselected miR-E + UltramiR set, showing a different nucleotide distribution due to the altered shRNA backbone. More shRNAs starting with a C were found to be potent in this set relative to TILE (p-value = 0.002, Fisher's exact test), indicating less restrictive sequence requirements when using the miR-E backbone.

The TILE dataset is unbiased as it was generated by completely tiling nine genes.

However, this design strategy produces a low fraction of potent shRNAs. To reduce costs and increase the ratio of potent shRNAs tested, subsequent screens only assessed shRNAs predicted to be efficient by various *in silico* methods^{11,12}. The M1, miR-E, and UltramiR datasets are based on preselected input libraries showing biased coverage of the sequence space and divergence in the nucleotide composition of potent shRNAs relative to the TILE dataset (**Figure 2.2**). When combined with the unbiased TILE set, these datasets comprehensively sample the distributions of features of non-functional and functional shRNAs. Effective integration of all sets is thus crucial for efficient miR-E shRNA prediction.

Combining diverse data sets presents a machine-learning challenge. Our approach of using a sequential classifier stems from classification strategies used in face detection^{41,42}, where a first classifier evaluates simple face-like features to reject obvious non-faces and a second classifier evaluates refined features on retained

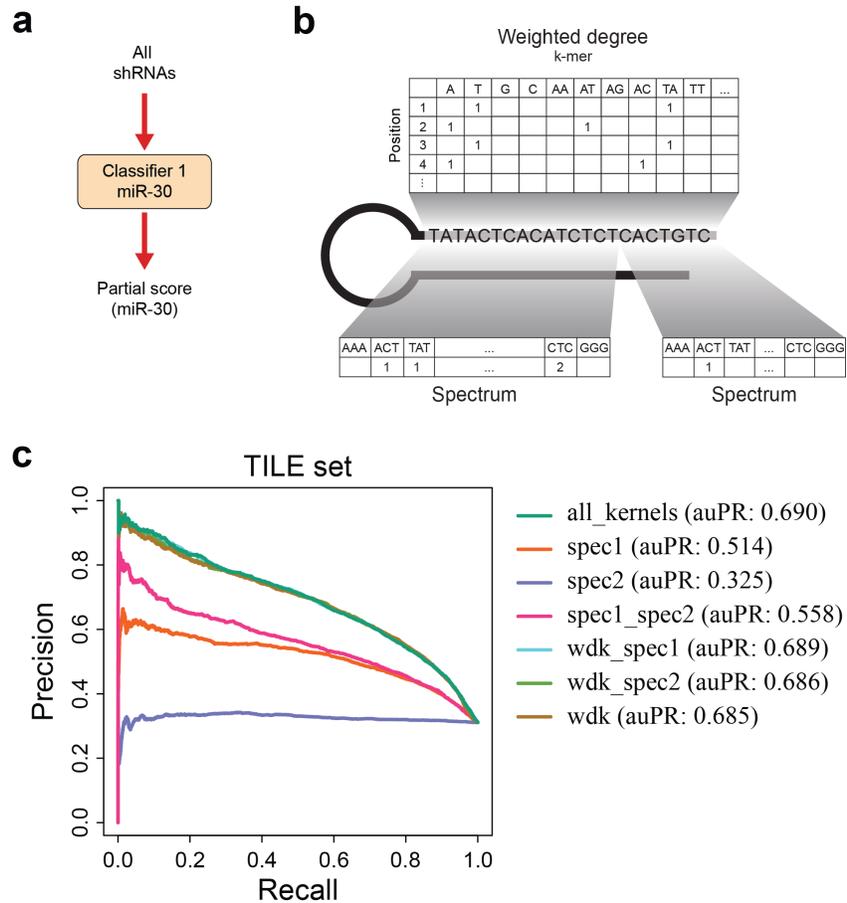


Figure 2.3: Kernel selection.

(a) Schematic of the first support vector machine (SVM) classifier that serves to eliminate non-functional sequences and prioritize shRNAs that are likely to be potent. **(b)** Schematic of the kernel representation used by SplashRNA. A weighted degree kernel is calculated across the entire guide sequence, while two spectrum kernels are calculated across nucleotides 1-15 and 16-22, respectively. **(c)** Testing of multiple kernel combinations in a leave-one-gene-out nested cross-validation setting on the TILE. All_kernels: wdk + spec1 + spec2. Spec1: spectrum kernel over positions 1-15. Spec2: spectrum kernel over positions 16-22. Spec1_spec2: spec1 + spec2. Wdk: weighted degree kernel over positions 1-22. Wdk_spec1: wdk + spec1. Wdk_spec2: wdk + spec2.

potential faces. Similarly, SplashRNA contains a sequence of two SVM classifiers trained on miR-30 and miR-E data. The miR-30 classifier evaluates shRNA sequence features to reject obvious non-functional shRNAs, whereas the miR-E classifier evaluates refined sequence features for retained, potentially potent shRNAs (**Figure 2.1c**). Each classifier is composed of a combination of k-mer feature representations^{43,44}. To capture AU content and position-specific k-mer features¹⁰, we represented an shRNA as a sum of a spectrum kernel on sequence positions 1–15, a spectrum kernel on sequence positions 16–22 and a weighted degree kernel on the entire sequence (**Figure 2.3b**). We found that this kernel combination yields the best performance (**Figure 2.3c**).

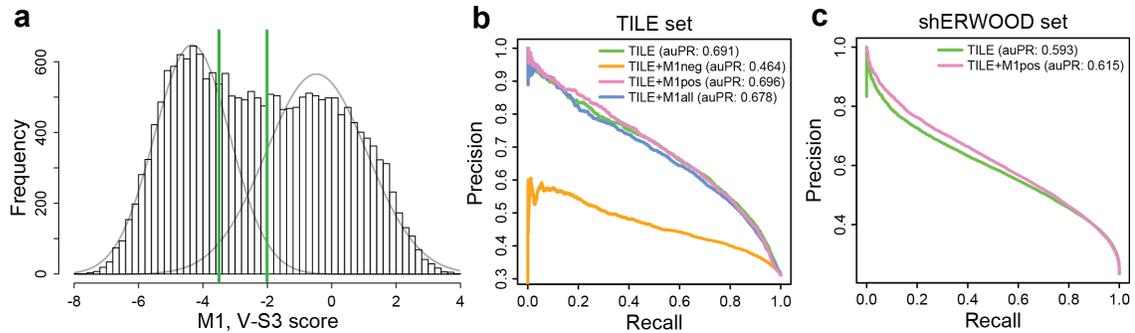


Figure 2.4: Incorporation of M1 dataset to generate a miR-30 classifier.

(a) M1 potency score distribution. Cutoffs (green lines) were calculated by fitting Gaussian distributions to the modes and setting thresholds at 5% FPR and 5% FNR. (b) Incorporation of M1 positives, negatives or both into the TILE training set was tested in a nested leave-one-gene-out cross-validation setting. Inclusion of M1 negatives deteriorated performance on the TILE dataset, whereas inclusion of the M1 positives alone improved performance. Note: TILE+M1pos = Splash_{miR-30}, the miR-30 classifier. (c) Incorporation of M1 positives into the TILE training set improved performance on the external shERWOOD miR-30 dataset.

Initially, we trained the miR-30 classifier on the combined positives and negatives from the TILE and M1 sets (**Table 2.1**). This yielded a classifier that scored well in validation tests but was outperformed by one trained on TILE alone (**Figure**

2.4b). We found that due to the biased selection strategy used to design the M1 screen, the potency of the negative shRNAs in the M1 screen were still more potent than the negative shRNAs from the TILE screen. Therefore, combining the M1 negatives with the TILE negatives degraded the performance of the model as they lowered the relative importance of the unbiased TILE negatives. Consequently, our best miR-30 classifier ($\text{Splash}_{\text{miR-30}}$) was obtained by training on a combined data set of TILE and M1 positives only (**Figure 2.4b,c**). The miR-E classifier ($\text{Splash}_{\text{miR-E}}$) was trained on the combined miR-E and UltramiR data sets using the same kernel combination. For the final SplashRNA predictor, $\text{Splash}_{\text{miR-30}}$ and $\text{Splash}_{\text{miR-E}}$ were combined by tuning two hyperparameters: theta (θ , a threshold above which predictions are passed to the second classifier) and alpha (α , the relative weighting of the scores from the two classifiers for predictions evaluated by the second classifier; **Figure 2.1c, Equation 2.1**). By calculating the precision-recall trade-off between the two classifiers, we chose values for theta and alpha that maintained the high performance of the first classifier while also predicting well on miR-E data (**Figure 2.5a-c**). Note that the performance of a sequential classifier equals or exceeds that of a linear combination since one can set the threshold to a large enough value such that all examples are evaluated by both classifiers (**Figure 2.5a**).

2.3 Training the component SplashRNA classifiers

When fitting the regularization parameter C for our miR-30 SVM, we used leave-one-gene-out nested cross-validation. We grouped shRNAs from the TILE miR-30 data set by target gene into outer-folds. Then for each outer fold, we held out shRNAs targeting one gene and optimized the parameter C on the shRNAs targeting the remaining genes through ten-fold cross-validation. The M1 positive set was added to all training sets but

was not used for selection of C or for validation. Performance on the TILE set is reported on the outer held-out genes (**Figure 2.4b**). We trained our final classifier with the parameter setting $C = 15$ using all the TILE and M1 positive shRNAs. This classifier was used to predict on all other data sets.

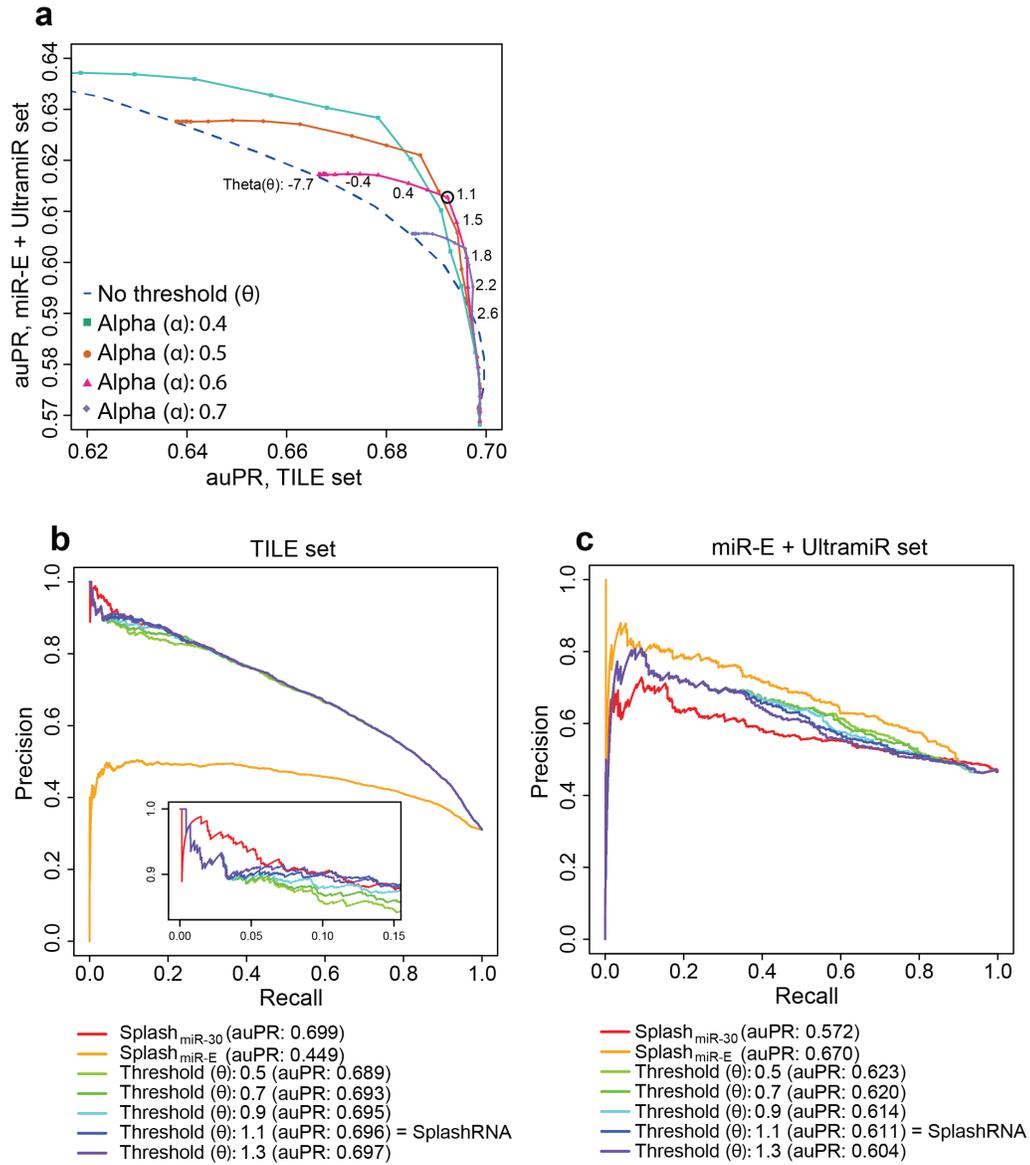


Figure 2.5: Calibration of the sequential SVM classifier SplashRNA.

(a) Precision-recall trade-off between the two classifiers Splash_{miR-30} and Splash_{miR-E}. Selection of alpha and theta hyperparameters leads to varied performance (area under the precision-recall curve, auPR) on the TILE miR-30 (x-axis) and miR-E + UltramiR (y-axis) sets. Each line represents a setting of alpha; points on the line represent distinct theta values. The circle indicates the alpha and theta choices for the final sequential classifier (SplashRNA, alpha = 0.6, theta = 1.1). Dotted line: performance of the convex linear classifier without a threshold at every alpha. (b) Performance on the TILE set, varying the value for theta with alpha set to 0.6. The insert shows the first 15% of the precision-recall. (c) Performance on the miR-E + UltramiR set, varying the value for theta with alpha set to 0.6.

Similar to the training regimen for the miR-30 SVM, we used nested ten-fold cross-validation to fit the C parameter for our miR-E SVM. We did not use a leave-one-gene-out strategy however due to the lower number of shRNAs targeting each gene. Instead, the miR-E and UltramiR sets were combined and split into ten outer folds. Within each of these folds, ten-fold cross validation was performed to determine the optimal C parameter for that fold. Performance on the miR-E and UltramiR sets is reported on the outer held-out folds (**Figure 2.5c**). We trained our final classifier with the parameter setting $C = 15$ using all the miR-E and UltramiR data. This classifier was used to predict on all other data sets.

In order to calculate the final SplashRNA score, the potency scores for all shRNA are first calculated using the miR-30 classifier. If the score does not exceed the threshold θ , this partial score is the final score for the shRNA. If the score does exceed the threshold, the final score is a weighted combination of the predicted scores from the miR-30 and miR-E classifiers.

$$Finalscore(x) = \begin{cases} \alpha SVM_{mir30}(x) & \text{if } \alpha SVM_{mir30}(x) < \theta \\ \alpha SVM_{mir30}(x) + (1 - \alpha) SVM_{miRE}(x) & \text{if } \alpha SVM_{mir30}(x) \geq \theta \end{cases}$$

2.4 SplashRNA outperforms existing shRNA prediction methods

When tested on miR-30 (**Figure 2.6a-c**) and miR-E (**Figure 2.6d**) data sets, SplashRNA clearly outperformed DSIR¹⁹, the current reference algorithm in the field (originally developed for siRNA design). The first classifier alone, Splash_{miR-30} (auPR: 0.615), shows the best performance. SplashRNA (auPR: 0.506) compromises slightly on miR-30 data to increase prediction accuracy on miR-E shRNAs, while still outperforming three other si/shRNA prediction tools (DSIR, seqScore, miR_Scan).

SplashRNA also outperformed the miR-30-based shERWOOD algorithm on the UltramiR set (**Figure 2.7a**), compared to its published maximum performance¹². We also observed the high performance of SplashRNA in two large-scale biological RNAi screens^{45,46} run with shRNAs functionally equivalent to miR-E (**Figure 2.7b,e**)⁴⁷, each of which tested ~25 preselected shRNAs per gene. In both screens, SplashRNA was able to retrospectively predict which shRNAs were potent and thus were enriched or depleted in the positive or negative selection screen, respectively. The positive selection data was selected from a large-scale pooled toxin sensitivity RNAi screen. Genes conveying sensitivity or resistance to the toxin were knocked down and the enrichment or depletion of shRNAs targeting each gene was measured. Similarly, the negative selection screen identified essential genes in K562 cells by measuring cell growth rate after knocking down genes. For each of the top 20 sensitivity genes and top 50 essential genes, all shRNA prediction algorithms selected their top and bottom five sequences and the log₂ fold changes for the selected shRNA were compared. SplashRNA was the only algorithm to achieve significant discrimination in the fold changes between the top and bottom predictions at $p < 0.01$ ($p = 4.8e-4$, one-sided Wilcoxon rank sum test) in the sensitivity gene study and achieved the most significant discrimination in the fold changes in the essential genes study ($p = 1.8e-11$, one-sided Wilcoxon rank sum test). Of note, SplashRNA also outperformed the other algorithms when selecting smaller or larger numbers of top sensitivity or essential genes from the screens (data not shown).

SplashRNA predictions also showed equally good or better accuracy compared to larger sets of preselected shRNAs when tested on a subset of the negative-selection screen that included only a previously established set of ‘gold-standard’ essential genes^{46,48}. The top ten SplashRNA predictions identified true positives significantly better than the bottom ten ($P < 0.001$, empirical permutation test), minimizing off-target hit

identification (**Figure 2.7d**). This indicates that the high performance of SplashRNA allows for fewer shRNAs to be tested per gene, decreasing the false discovery rate as well as the cost of the screen.

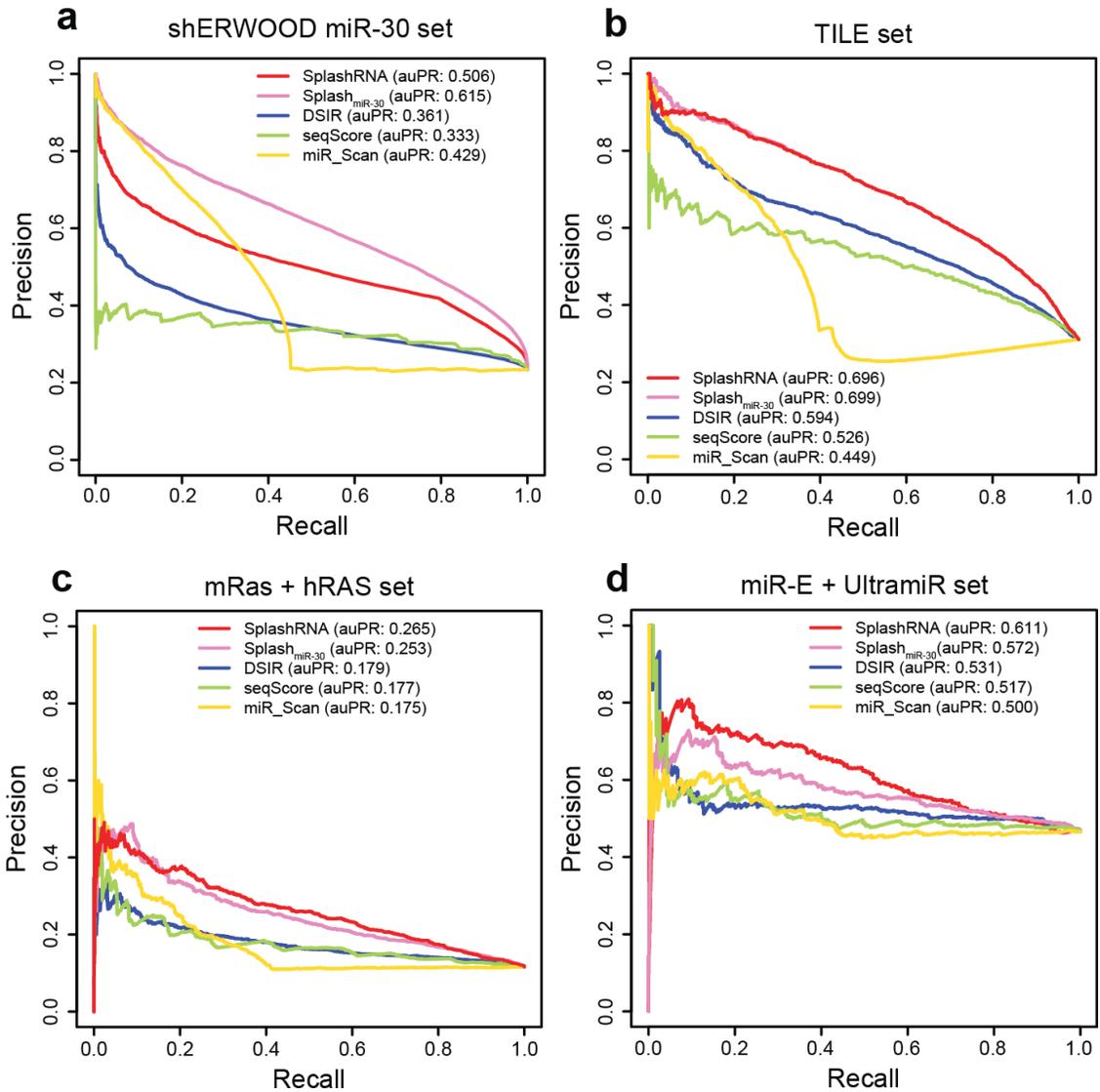
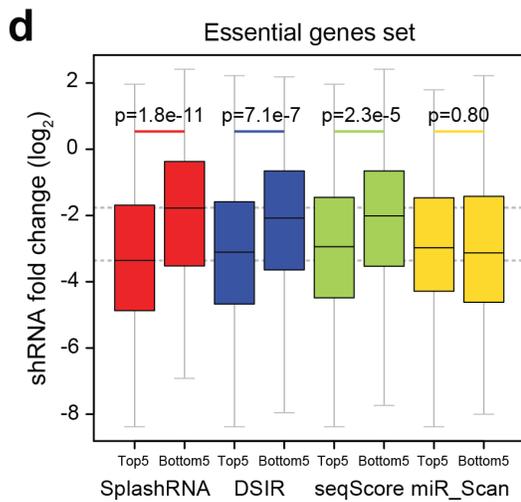
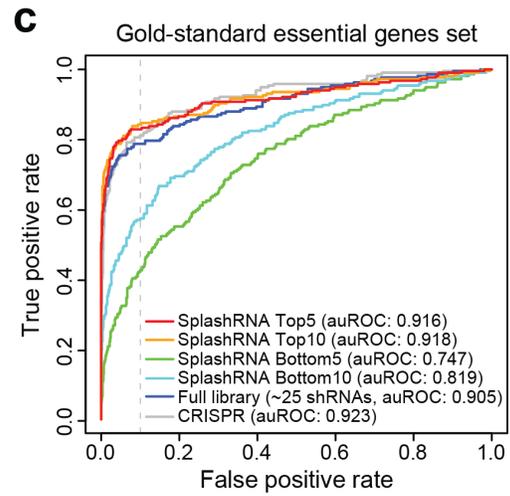
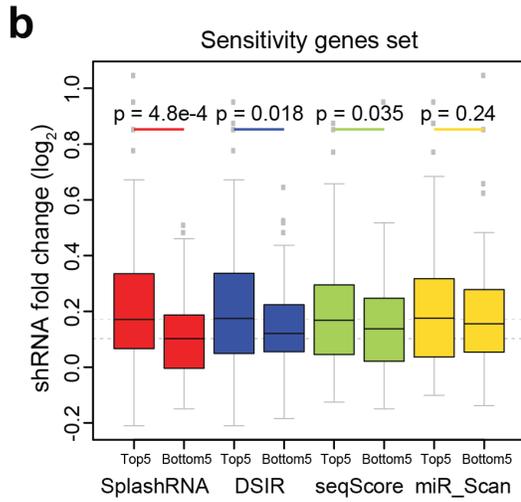
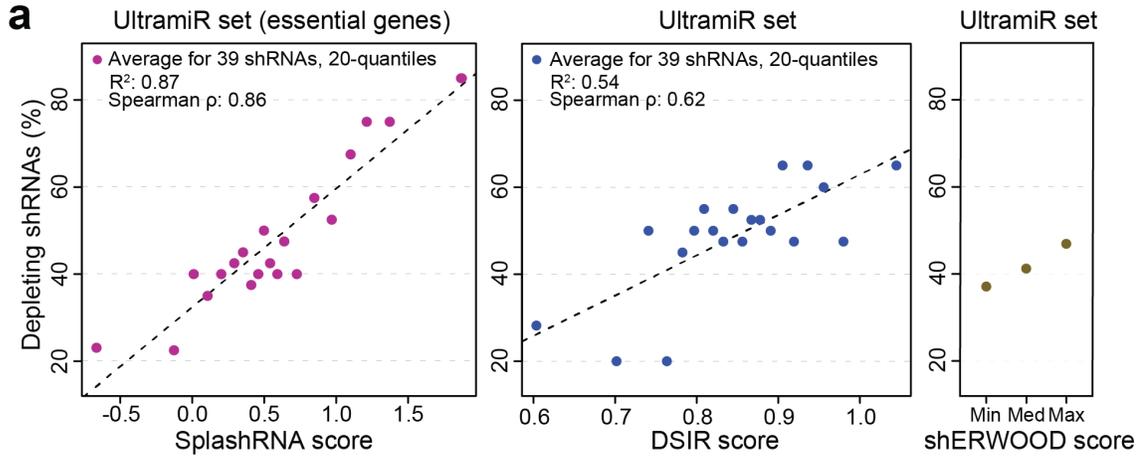


Figure 2.6: Performance of various shRNA prediction algorithms.

(a) Precision-recall curves of SplashRNA performance on the external shERWOOD miR-30 dataset. (b) Precision-recall curves on the TILE dataset, comparing leave-one-gene-out nested cross-validation predictions from SplashRNA (auPR: 0.696) and Splash_{miR-30} (auPR: 0.699) against the alternative prediction tools DSIR (auPR: 0.594), seqScore (auPR: 0.526) and miR_Scan (auPR: 0.449). (c) Prediction performance comparison of the indicated algorithms on the external mRas + hRAS Sensor dataset (Table 2.1). (d) SplashRNA performance on miR-E data. SplashRNA (auPR: 0.611) clearly outperforms the miR-30 classifier alone (auPR: 0.572) as well as three other prediction tools.

Figure 2.7: SplashRNA performance on *in vivo* screens.

(a) SplashRNA and DSIR based re-ranking of shERWOOD selected UltramiR shRNAs targeting essential genes tested in a cell viability screen. X-axis: mean SplashRNA or DSIR score for equally sized groups (purple and blue dots, 20 groups) of 39 shRNAs each. Y-axis: Percent of shRNAs in each group that were potent. Right panel: published minimum, median, and maximum (Max) shERWOOD performance (green-brown dots). **(b)** Log₂ fold changes in the top and bottom 5 retrospective potency shRNA predictions from SplashRNA and competing algorithms for the top 20 most sensitizing genes from a large-scale toxin RNAi screen. **(c)** ROC curve comparing algorithms' ability to identify "gold-standard" essential genes. The dashed line represents the 10% false positive rate (FPR) threshold. **(d)** Log₂ fold changes in top and bottom 5 shRNA predictions for the top 50 most essential genes from a large-scale essential genes RNAi screen.



2.5 Targeting the relevant transcript space

Robust shRNA prediction starts with the selection of the right transcript region. Analyses of unbiased TILE data showed that efficient shRNAs are more prevalent in 3' UTRs compared to coding sequences and 5' UTRs (**Figure 2.8a**), likely due to relatively high AU content (**Figure 2.8b–d**)¹⁰. Whereas 3' UTRs often present ample design space because of their lengths, when validating top predictions in mouse fibroblasts, many shRNAs targeting the distal end of Pten resulted in minimal or no protein knockdown (**Figure 2.8e, Table 2.2**). Inspection of the Pten mRNA (NCBI, NM_008960) revealed that all these low-potency shRNAs targeted regions past polyadenylation signals (PASs), the use of which lead to shorter transcript variants⁴⁹ lacking the respective target sites (**Table 2.2**). Hence, to eliminate alternative cleavage and polyadenylation (ApA) as a source of non-functional shRNAs, we used PAS atlases^{50,51} to annotate the human and mouse reference transcriptomes (NCBI) and discard 3' UTR portions that may be absent due to cleavage at a signal early in the 3' UTR. Similarly, we report predictions only on the intersection of all transcript variants for each gene and filter multi-matching sequences.

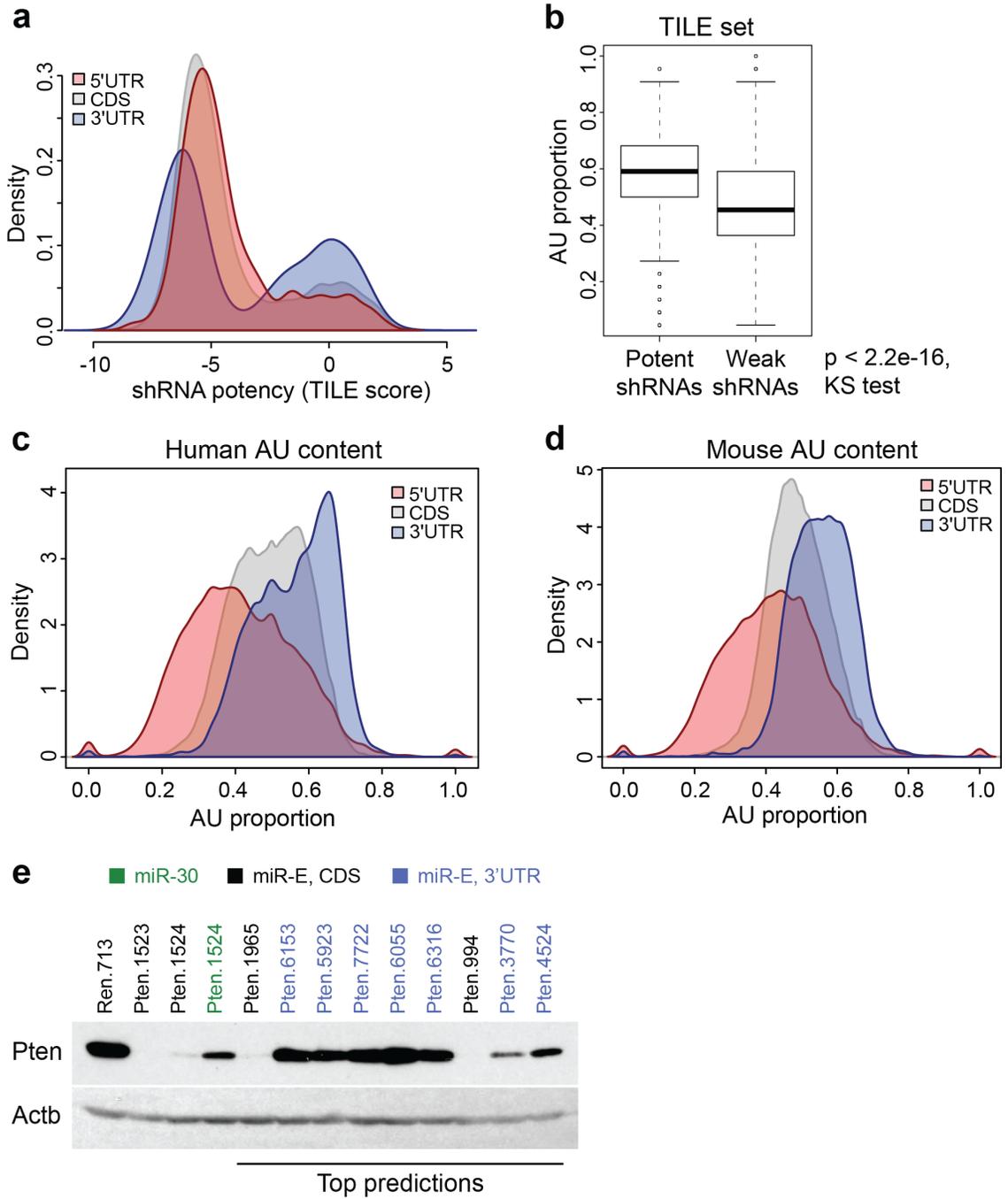
Similarly, alternative splicing can lead to sequences being included or excluded in some gene transcripts. In order to ensure that a predicted shRNA will target all isoforms of a gene, we predict only on sequences that are present in all isoforms of a gene. Additionally, we do not predict shRNAs that span splice junctions to avoid mis-targeting of the predicted shRNA due to variations in splicing.

Table 2.2: Polyadenylation sites in Pten gene affect observed shRNA knockdown. Start and End indicate shRNA site in base pairs on the 19th chromosome of the mm9 build of the mouse genome. KD: qualitative degree knockdown observed in immunoblotting analyses of NIH/3T3s. PAS: previously published locations of polyadenylation sites, 19th chromosome, mm9, identified in NIH/3T3 and mouse ES cells by 3P-seq. 2P-Seq: quantification of transcript expression levels measured by 2P-Seq. All shRNAs and PASs are ordered according to their position along the mouse genome (mm9).

shRNA	Start	End	KD	PAS	2P-Seq	Comment
Pten.994	32850552	32850573	+++			
Pten.1523	32889925	32889946	+++			
Pten.1524	32889926	32889947	+++			
Pten.1965	32894404	32894425	+++			
				32894814	184.0	Major PAS
Pten.3770	32896209	32896230	++			
Pten.4524	32896963	32896984	+			
				32897818	47.0	Minor PAS
Pten.5923	32898362	32898383	-			
Pten.6055	32898494	32898515	-			
Pten.6153	32898592	32898613	-			
Pten.6316	32898755	32898776	-			
Pten.7722	32900161	32900182	-			
				32900648	49.5	Poly(A)

Figure 2.8: Transcript selection.

(a) Distribution of shRNA potency in functionally distinct transcript regions. Shown is the potency distribution of shRNAs in the unbiased TILE dataset that target the 5'UTR, CDS or 3'UTR. Since these shRNAs were evaluated using the Sensor assay, their targets are not subject to alternative cleavage and polyadenylation (APA) and/or splicing events. **(b)** A/U content of potent and weak miR-30 shRNAs from the unbiased TILE set. Potent shRNAs tend to have a higher proportion of A/U nucleotides ($p < 2.2e-16$, two-sided Kolmogorov-Smirnov test). **(c)** A/U content of functionally distinct transcript regions in the human genome. Shown are the A/U densities in 5'UTR, CDS and 3'UTR. **(d)** A/U content in mouse transcripts. **(e)** Alternative cleavage and polyadenylation (ApA) prevents potent shRNAs from inhibiting their putative target gene. Immunoblotting of Pten in NIH/3T3s transduced at single-copy with LEPG expressing the indicated shRNAs. Nine top predictions targeting the CDS or the 3'UTR after early ApA sites were compared alongside controls for their ability to suppress mouse *Pten*. *Actb* was used as loading control.



2.6 *In vivo* validation of *de novo* predicted shRNAs

Testing an extensive set of individual *de novo* predictions targeting *Pten*, *Bap1*, *Pbrm1*, *Rela*, *Bcl2l11*, *Axin1*, *NF2* and *Cd9* under single-copy conditions¹³ by conventional western blot analysis (**Figure 2.9a-f,h**) or flow-cytometry based immunofluorescence of surface proteins (**Figure 2.10a**), we found that protein knockdown levels were very high: 91% of predictions (41/45) with a SplashRNA score of greater than 1 showed 85% or higher protein knockdown (**Figure 2.10b**). Even in the case of human NF2, a gene with nine annotated transcript variants that share only 198 nucleotides (excluding the 5' UTR, **Figure 2.9g**), the top eight SplashRNA predictions triggered 77–96% (median 89%) protein suppression under single-copy conditions (**Figure 2.9h**). Additionally, Cd9 knockdown analyses in mouse fibroblasts showed that SplashRNA clearly outperforms DSIR in *de novo* prediction and achieves near knockout levels comparable to CRISPR–Cas9 (**Figure 2.10a**).

Extrapolating beyond the tested shRNAs, we calculated the proportion of genes for which SplashRNA would find at least five shRNAs above a given threshold (**Figure 2.10b**). After shortening of transcripts due to ApA and considering only the intersection of all transcript variants per gene, we found that 87% of mouse genes and 81% of human genes have at least five shRNAs with SplashRNA scores above 1, corresponding to an 80% probability (e.g., four out of five shRNAs) of more than 85% knockdown at single-copy (**Figure 2.10c**).

Figure 2.9: Western blot validation of *de novo* SplashRNA predictions.

All shRNAs were expressed using LEPG at single-copy conditions. β -Actin (Actb, ACTB) was used for normalization. Long (top) and short (bottom) exposures are shown. Immunoblotting of (a) Pten, (b) Bap1, (c) Pbrm1, (d) Rela, (e) Bcl2l11, (f) Axin1 in NIH/3T3s. C, miR-30 and miR-E control shRNAs. (g) Graphical depiction of human NF2 transcript variants. NF2 has nine variants with an intersection of only 198 nucleotides, excluding the 5'UTR, rendering the prediction task especially difficult due to limited sequence space. (h) Predicting miR-E shRNAs for short transcripts. Immunoblotting of NF2 in A375s transduced with the indicated shRNAs targeting all nine NF2 variants.

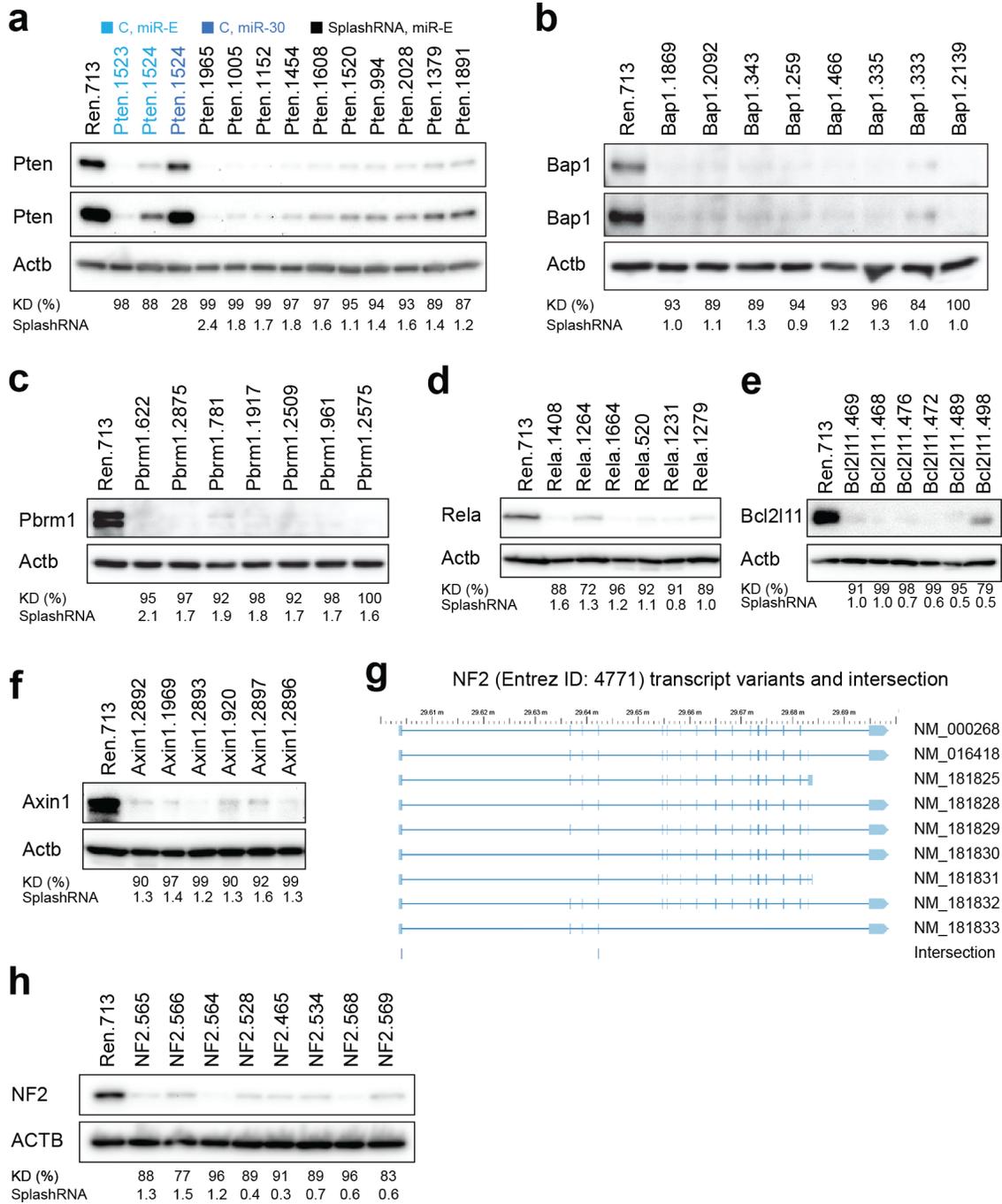
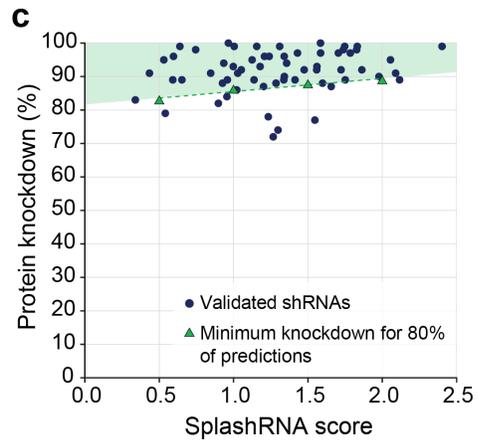
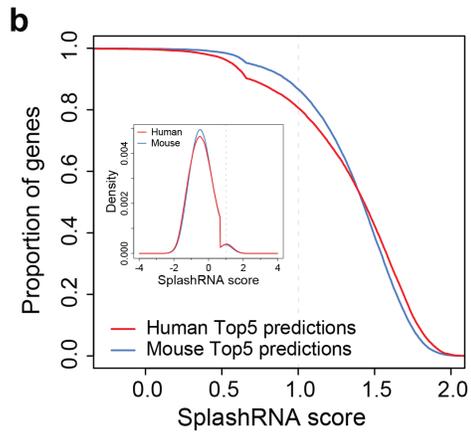
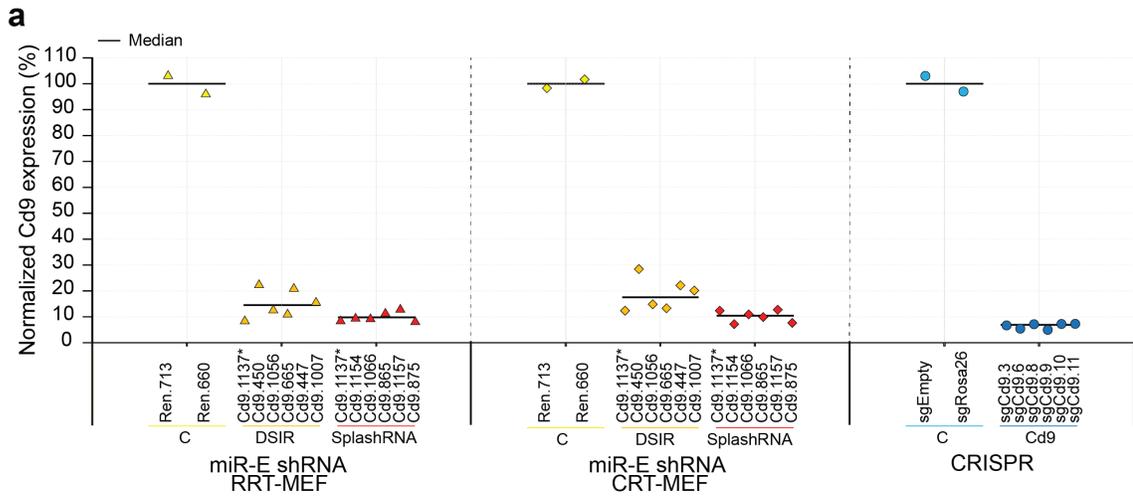


Figure 2.10: SplashRNA comparison to CRISPR-Cas9

(a) Comparison of SplashRNA and DSIR predictions against CRISPR-Cas9 mediated suppression of Cd9 in mouse embryonic fibroblasts (MEFs). Shown are normalized (relative to the indicated controls) median anti-Cd9-APC fluorescence intensities of RRT-MEFs and CRT-MEFs expressing the indicated shRNAs or sgRNAs (**Methods**). The six top-scoring predictions from DSIR + Sensor rules (DSIR) or SplashRNA (ordered according to their respective scores) were compared to six sgRNA sequences. *, Cd9.1137 is the top prediction of both algorithms and was plotted twice for clarity. While DSIR predictions triggered Cd9 knockdown with variable efficacy, SplashRNA predictions consistently induce strong Cd9 suppression, closely approaching knockout conditions. **(b)** Score distribution of fifth highest SplashRNA prediction for all human and mouse genes, indicating the proportion of genes with 5 predictions above a given score. Predictions were run only on the intersection of all transcript variants per gene and after shortening of transcripts to the most proximal PAS. The inset shows the score distribution of all human and mouse SplashRNA predictions. The kink in the curves represents the transition from Splash_{miR-30} to combined SplashRNA scores. At least 80% of genes have five shRNAs with prediction scores above 1 (dotted line). **(c)** Transfer function of SplashRNA score versus protein knockdown for all 62 *de novo* predicted shRNAs validated by immunofluorescence. Green triangles indicate the minimum knockdown for 80% of the predictions for a given SplashRNA score bin. Bins were defined to have a width of 0.5 with the leftmost bin starting at 0.25. For the bin centered on SplashRNA score = 1, 80% of predictions showed at least 86% protein knockdown. The expected knockdown for the top 80% of predictions (e.g. 4/5 shRNAs) increases with the SplashRNA score. Together, 91% of predictions with a SplashRNA score >1 showed more than 85% protein knockdown.



2.7 Minimization of off-target effects

Though RNAi triggers can be expressed as simple stem-loop shRNAs from RNA polymerase III (Pol-III) promoters in mammalian cells, such strategies can lead to off-target effects associated with high shRNA expression levels³⁵, likely due to saturation of the endogenous microRNA machinery⁵². Many Pol-III-based systems also suffer from inaccurate processing of precursor molecules³⁴, yielding undesired mature small RNAs. In contrast, use of microRNA-embedded shRNAs expressed from RNA polymerase II (Pol-II) promoters results in accurate processing^{39,40} and can alleviate the toxic side effects^{36,37,53}, especially when used at single genomic integration (single-copy)¹¹. Notably, highly potent miR-30-based shRNAs expressed at single-copy show the same low levels or absence of off-target effects as analogous weak and non-functional sequences¹¹. Hence, to develop an improved shRNA prediction algorithm, we focused on the optimized miR-E system that is based on the endogenous human MIR30A¹³.

Here, to determine the extent of sequence-based off-target effects we applied the GESS (Genome-wise Enrichment of Seed Sequence) algorithm⁵⁴ to shRNAs validated by immunoblotting (**Figure 9a-f,h**), and to previously reported Sensor assay and gene expression microarray results^{10,11}. We tested if potent shRNAs have more off-target effects than their weaker counterparts and if these targets have common sequences. First, to investigate sequence-based off-target effects, we analyzed RNA expression microarray data from Trp53^{-/-} MEF cells infected at single or high copy with one of six Trp53 shRNAs¹¹. Repetition of the published differential expression analysis found zero differentially expressed genes in the single-copy transfection setting relative to control experiments for either potent or weak shRNAs. In the high-copy transfection setting, 702 genes were up-regulated and 326 genes were down-regulated in the cells with

potent shRNA with respect to control experiments (FDR < 0.05). Additionally, 2,437 genes were up-regulated and 1,731 genes were down-regulated in cells transfected with weak shRNA relative to their controls. Therefore, potent shRNAs in this setting did not induce more gene expression changes than weak shRNAs but high-copy transfection did induce more off-target effects. Furthermore, the high-copy transfections of both the potent and weak shRNAs resulted in near identical lists of differentially expressed genes: 702 of 702 genes were significantly up-regulated in both lists and 324 of 326 genes were significantly down-regulated in both lists. These intersections significantly overlapped (up-regulated: $P < 2.2 \times 10^{-16}$, down-regulated: $P < 2.2 \times 10^{-16}$, Fisher's exact test), indicating that the main changes in gene expression are similar regardless of potency or shRNA sequence composition.

Second, we applied the GESS algorithm⁵⁴ to our validation shRNAs that were quantified by immunoblotting to determine potential sequence-based off-target effects in our current experiments. We attributed our shRNAs to three categories based on western blot knockdown: Low (less than 80% knockdown), Mid (between 80% and 95% knockdown), High (95% knockdown or greater). For each gene and potency-level group, we ran GESS and found the genes that were potentially targeted by all the shRNAs in the group. We found no statistically significant off-targeted genes by GESS (FDR < 0.1). We also tested if the level of potency is associated with the number of potential off-target genes as measured by the number of perfect 7-mer seed matches (nucleotides 2–8). Grouping shRNAs into three groups by percent knockdown, High: >95%, Medium: 90–95%, and Low: 80–90%, and testing for a significant difference in the number of gene seed matches found no statistically significant difference between any pair of groups ($P = 0.74, 0.53, \text{ and } 0.73$ for Low vs. Medium, Low vs. High, and Medium vs. High, respectively).

Third, we calculated all perfect 22-mer multi-mapping matches transcriptome-wide, since perfect matching of an shRNA to several genes would be highly undesirable. Consequently, we incorporated an additional feature into the SplashRNA algorithm and web site that alerts the user if a predicted hairpin perfectly matches multiple genes in the human or mouse transcriptomes (hg38, mm10).

2.8 Discussion

Building on our Sensor assay and the optimized miR-E backbone, here we have established a robust algorithm to predict ultra-potent microRNA-based shRNAs targeting nearly any gene. SplashRNA is able to accurately predict the potency of independently validated and novel shRNAs and outperforms existing algorithms. Our sequential predictor approach facilitates the integration of biased and unbiased data sets and can serve as a blueprint for other prediction problems. An open source implementation of SplashRNA is accessible at <http://splashrna.mskcc.org>. The website can directly predict on custom sequences or mouse and human Entrez Gene IDs. When using gene IDs, the tool integrates cleavage and poly-adenylation signal annotations and calculates the intersection of all transcript variants to predict only on constitutive sequence regions. To facilitate the use of SplashRNA, all transcript annotations and predicted shRNAs are graphically displayed online and can be downloaded in batch.

CHAPTER 3

CHROMATIN STATES DEFINE TUMOR-SPECIFIC T CELL DYSFUNCTION AND REPROGRAMMING

Portions of this chapter first appeared in Philip et. al.⁵⁵ and were written in collaboration with Mary Philip, Andrea Schietinger, and Christina Leslie.

3.1 Introduction

Tumor-specific CD8 T cells (TST) are often found within solid tumors, but tumors progress despite their presence, suggesting that these TST are dysfunctional⁵⁶. The clinical success of immune checkpoint blockade (for example, PD1/PDL1- and CTLA4-blocking antibodies) and adoptive T cell therapy in a subset of patients with cancer demonstrates the great potential of TST⁵⁷; however, important questions remain, including how to predict which patients will respond to therapy and precisely which TST mediate clinical responses⁵⁸⁻⁶⁰. Moreover, an unmet need is the development of interventions for tumors that are refractory to checkpoint blockade despite having ample TST infiltration.

We previously demonstrated that in the early stages of tumorigenesis, TST become non-responsive, exhibiting the phenotypic, functional, and transcriptional features of tumor-reactive tumor-infiltrating lymphocytes (TIL) from late-stage human solid tumors³⁰.

TST dysfunction is initially reversible but ultimately becomes irreversible, even after removal of dysfunctional T cells from the tumor microenvironment and multiple rounds of cell division³⁰. We hypothesized that this heritable, signal-independent dysfunctional state is epigenetically imprinted. The epigenetic programs that regulate normal differentiation of innate and adaptive lymphocytes have been described⁶¹⁻⁶⁴. However, the epigenetic programs regulating T cell differentiation and dysfunction in tumors are

not known. In this study, we used the assay for transposase-accessible chromatin with high-throughput sequencing (ATAC-seq)³² to assess genome-wide chromatin accessibility changes during T cell differentiation in tumors compared to acute infection.

3.2 CD8 T cell chromatin changes during infection

We transferred congenically marked naïve (N; CD44^{lo}CD62L^{hi}) TCR_{TAG} cells (specific for SV40 large T antigen epitope I (TAG))⁶⁵ from TCR_{TAG} transgenic mice into wild-type C57BL/6 mice, which were immunized one day later with a recombinant *Listeria monocytogenes* strain expressing TAG (*LmTAG*)^{30,66}. TCR_{TAG} cells were re-isolated, phenotypically and functionally characterized, and underwent ATAC-seq and RNA-seq at 5, 7 (effectors; E5, E7) and 60+ days (memory; M) after immunization (**Figure 1.4a**).

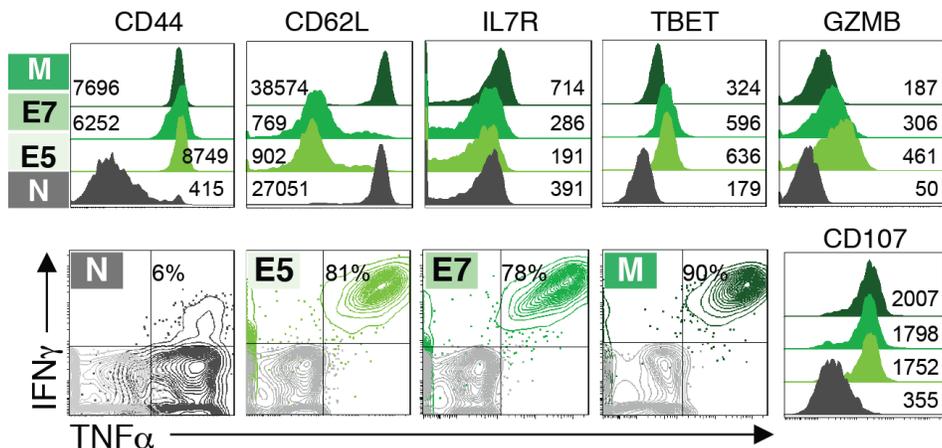


Figure 3.1: Flow cytometric analysis of characteristic markers after *LmTAG* stimulation.

Naïve TCR_{TAG} (N; Thy1.1⁺) were transferred into B6 (Thy1.2⁺) mice which were immunized with *LmTAG* one day later. At days 5, 7, and 60+ post *LmTAG*, effector (E5 and E7) and memory (M) T cells were isolated from spleens and assessed for phenotype and function. Flow cytometric analysis of CD44, CD62L, IL7R α , TBET, and GZMB expression *ex vivo* (upper panel), and intracellular IFN γ and TNF α production and CD107 expression after 4-hour *ex vivo* TAG peptide stimulation (lower panel). Flow plots are gated on CD8⁺ Thy1.1⁺ cells.

N, E5, E7, and M expressed characteristic activation, homing and cytokine receptors (CD44, CD62L, IL7R), transcription factors (TBET), cytotoxic molecules (GZMB, CD107), and pro-inflammatory cytokines (IFN γ , TNF α) (**Figure 3.1**).

Using DESeq2⁶⁷ to assess differential chromatin accessibility, we found that substantial chromatin remodeling occurred as cells differentiated from the N to the

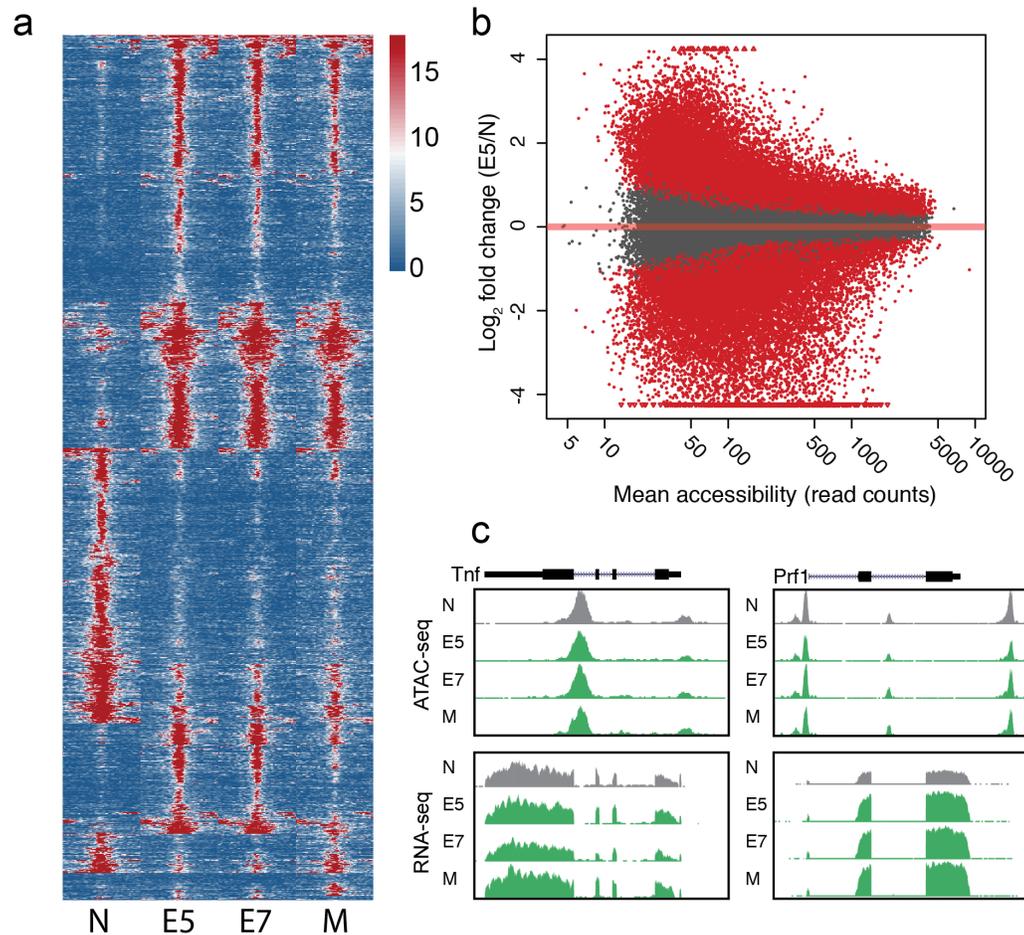


Figure 3.2: Chromatin accessibility in normal CD8 T cells

(a) Chromatin accessibility heat map grouped by differential accessibility patterns. Each row represents one of 8,654 selected peaks (significantly differentially accessible between at least one sequential cell comparison; FDR < 0.05, absolute log₂ fold change > 2). (b) MA plot, naive (N) to day 5 effectors (E5) transition, mean read counts for all atlas peaks versus log₂ ratios of peak accessibility. Significantly differentially accessible peaks are shown in red (FDR < 0.05). (c) ATAC-seq (top) and RNA-seq (bottom) signal profiles of *Prf1* and *Tnf* in naive, effector, and memory TCR_{TAG} cells during acute *Lm*TAG infection.

effector state (E5) (37,038 ATAC-seq peaks were differentially accessible during this transition), with substantially less remodeling from E5 to E7 and E7 to M (23 and 6,459 differentially accessible regions, respectively, FDR < 0.05, **Figure 3.2a,b**). In naïve cells, effector gene loci such as *Prfl* and *Tnf* shared highly accessible chromatin and basal transcriptional activity with effector and memory cells (**Figure 3.2c**), consistent with the presence of activation-associated histone marks previously shown at these loci in naïve T cells^{68,69}

We analyzed accessibility changes during the N to E5 transition in loci associated with early and late TCR-response genes, as defined by the Immunological Genome Project⁷⁰. Early-response genes showed many fewer accessibility changes compared to late-response genes ($p < 1e-16$), implying that they may not require

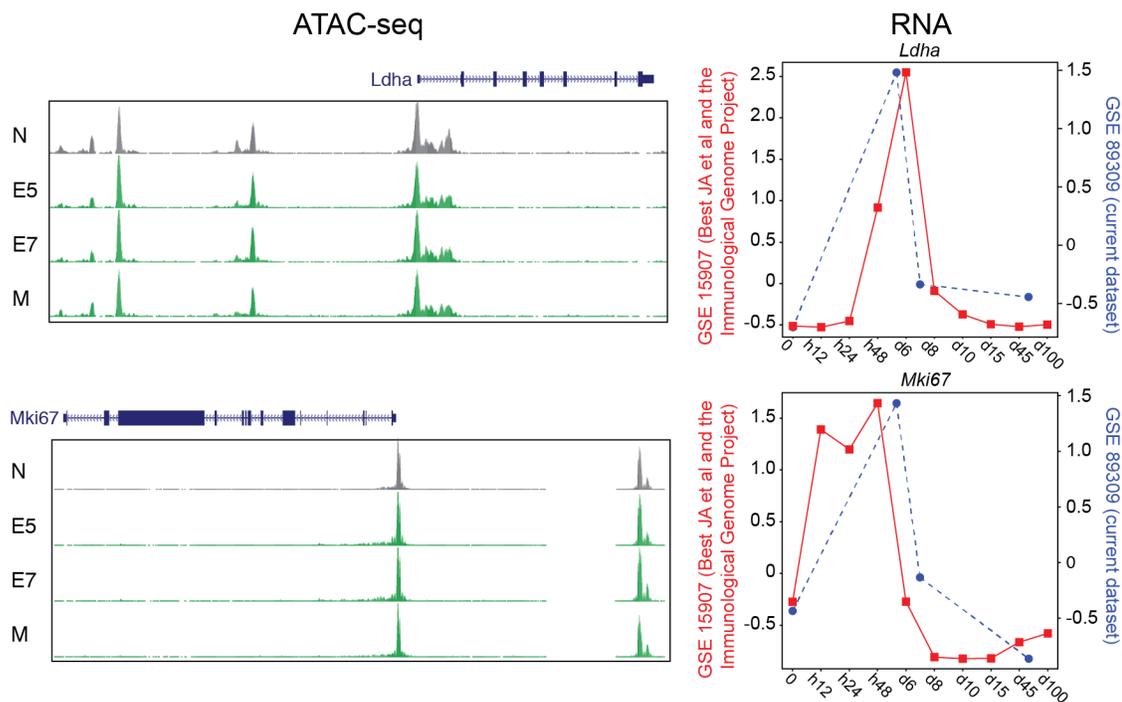


Figure 3.3: Activation of early response genes.

ATAC-seq signal profiles (left) and z-score normalized RNA expression (right) of the early response genes *Ldha* (top) and *Mki67* (bottom) in N, E5/E7, and M TCR_{TAG} cells during acute *Lm*TAG infection (blue line, current data set) overlaid with expression data from Immunological Genome Project (red line).

chromatin accessibility changes to activate expression. For example, *Ldha* (encoding LDHA, required for the metabolic shift to aerobic glycolysis and IFN γ production⁶⁹) and *Mki67* (encoding KI67, required for chromosome segregation during mitosis⁷¹) loci require no change in chromatin accessibility to be rapidly induced after TCR stimulation (**Figure 3.3**).

Memory T cells exhibit more rapid and robust effector function upon antigen re-encounter compared to naive T cells⁷². K-means clustering of RNA expression patterns

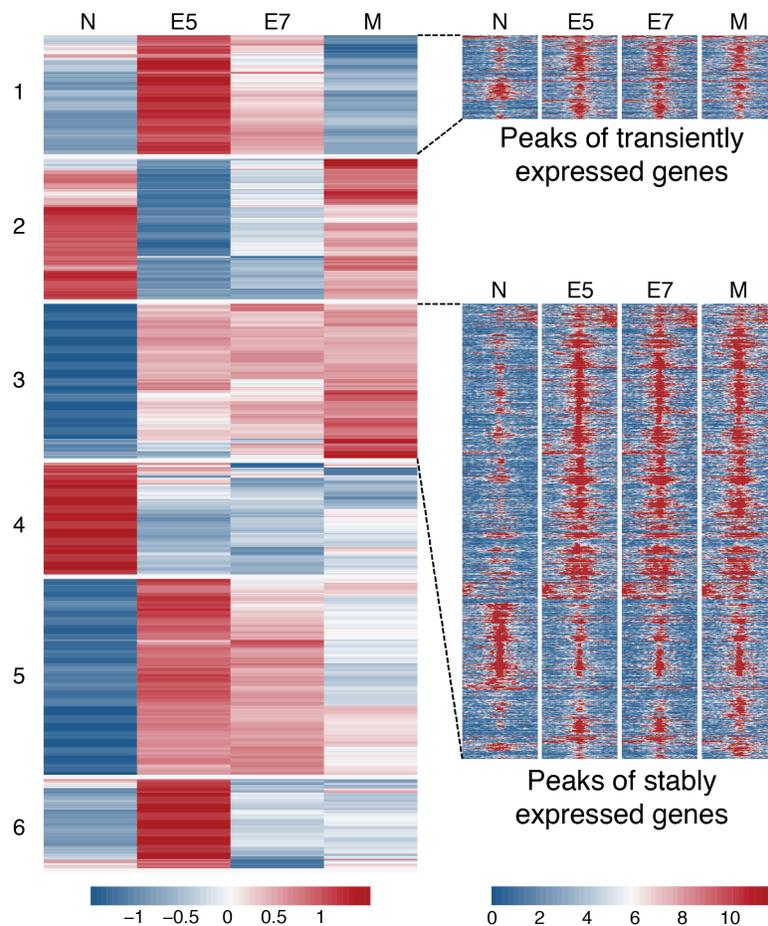


Figure 3.4: Linked k-means clustered RNA-seq and ATAC-seq heatmaps. Left, K-means clustered (K=6, row-normalized) RNA-seq data for 1,758 differentially expressed genes ($\log_2(\text{FC}) > 1$, $\text{FDR} < 0.05$, base mean \log_2 expression ≥ 10). Right, heat map of differentially accessible peaks ($\text{FDR} < 0.05$, $\log_2(\text{FC}) > 1$) associated with genes in K-means clusters 1 and 3. Right scale: z-scored Log_2 read counts, Left scale: Log_2 read counts per 20bp bin.

(Figure 3.4, left) revealed two trends: transient gene activation or down-regulation in E5/E7 but not M (clusters 1, 2, 5, 6), and stable gene activation or down-regulation in E5, E7, and M (clusters 3 and 4). Surprisingly, chromatin accessibility of loci identified as transient was largely similar in effector and memory cell states (Figure 3.4, right), indicating that memory cells are retaining an “effector-like” chromatin state. This retained chromatin state may permit basal transcription of certain effector genes (cluster 3) such as *Ifng*, whereas other genes are transcriptionally silent but poised for rapid re-expression upon TCR activation (cluster 1, *Gzma*) (Figure 3.4, right).

3.3 Chromatin state dynamics of TST dysfunction

We next assessed chromatin-state dynamics in TST over the course of tumorigenesis using the previously described tamoxifen-inducible, autochthonous liver cancer model (AST-Cre-ER^{T2}) in which TAG is a tumor-specific antigen³⁰. AST-Cre-ER^{T2} mice initially develop pre-malignant lesions which progress into hepatocellular carcinoma by

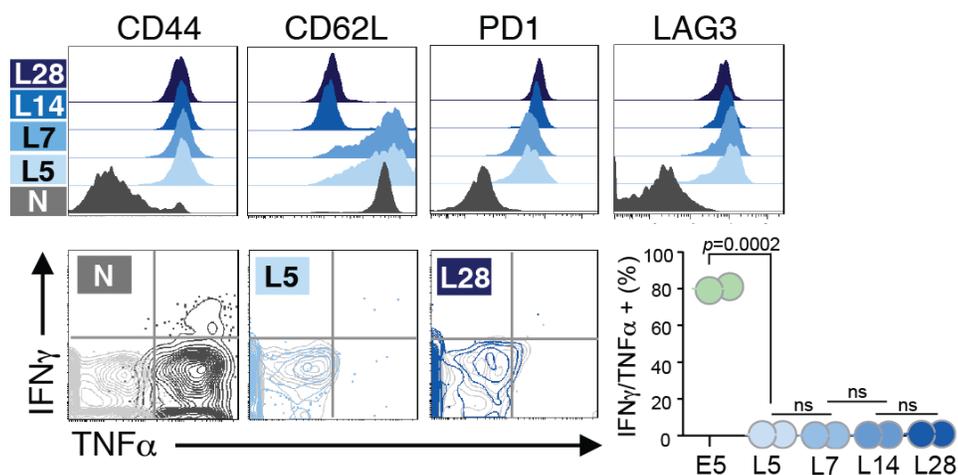


Figure 3.5: Flow cytometric analysis of inhibitory markers in TST.

Immunophenotype and cytokine production (grey, no peptide control) (n=8 total, n=2 per time point). Each symbol represents an individual mouse. *P=0.0002 (Student’s t-test); NS, not statistically significant.

day 60–90³⁰. We transferred congenically marked naive TCR_{TAG} cells (N, the same as N in **Figure 1.4a**) into AST-Cre-ER^{T2} mice one day before administration of tamoxifen and then analyzed TCR_{TAG} cells at different time points (**Figure 1.4b**). Liver-infiltrating TCR_{TAG} cells down-regulated CD62L, uniformly expressed activation markers CD44 and inhibitory receptors PD1 and LAG3, and failed to produce IFN γ or TNF α (**Figure 3.5**). Massive chromatin remodeling occurred by day 5: about 50,000 ATAC-seq regions were significantly differentially accessible between day 5 and naïve cells, followed by a second wave of remodeling between days 7 and 14, resulting in 25,000 differentially accessible regions (FDR<0.05, **Figure 3.6a**). Notably, after this second wave, few accessibility changes occurred, even after progression to established tumors at day 60+ (**Figure 3.6a,b**). Thus, TST differentiated through two discrete chromatin states: an initial state (L5, L7), and a later state established by day 14 and persisting thereafter. Many of the ATAC-seq peaks that were gained or lost during these transitions were in intronic and intergenic regions (potential enhancer peaks), whereas peaks present across all CD8 T cells were predominantly in promoter regions (**Figure 3.6c, bottom**); this pattern was also present in functional CD8 T cell differentiation (**Figure 3.6c, top**).

TCR_{TAG} cells in malignant lesions followed a distinct epigenetic trajectory compared to TCR_{TAG} cells in acute infection (L5 versus E5; **Figure 3.6b**). Many accessibility changes occurred exclusively in the transition from N to early dysfunctional (N to L5, 3641 opening, 3213 closing) or N to functional effector (N to E5, 9696 opening, 6271 closing). Two intergenic ATAC-seq peaks near the *Ifng* locus that opened during normal effector differentiation were inaccessible in dysfunctional TCR_{TAG} cells (**Figure 3.7, right**), indicating potential dysregulation of *Ifng*. Additionally, an intergenic peak 23.8 kb upstream of the PD1-encoding *Pdcd1* locus was accessible in all tumor-associated T cells (L5 to L60+), but not in N, E5/E7, or M (**Figure 3.7, left**). This

region was also described in exhausted T cells in chronic viral infection⁷³⁻⁷⁵ indicating a potential mechanistic similarity between tumor-associated T cell dysfunction and exhaustion associated with chronic viral infection.

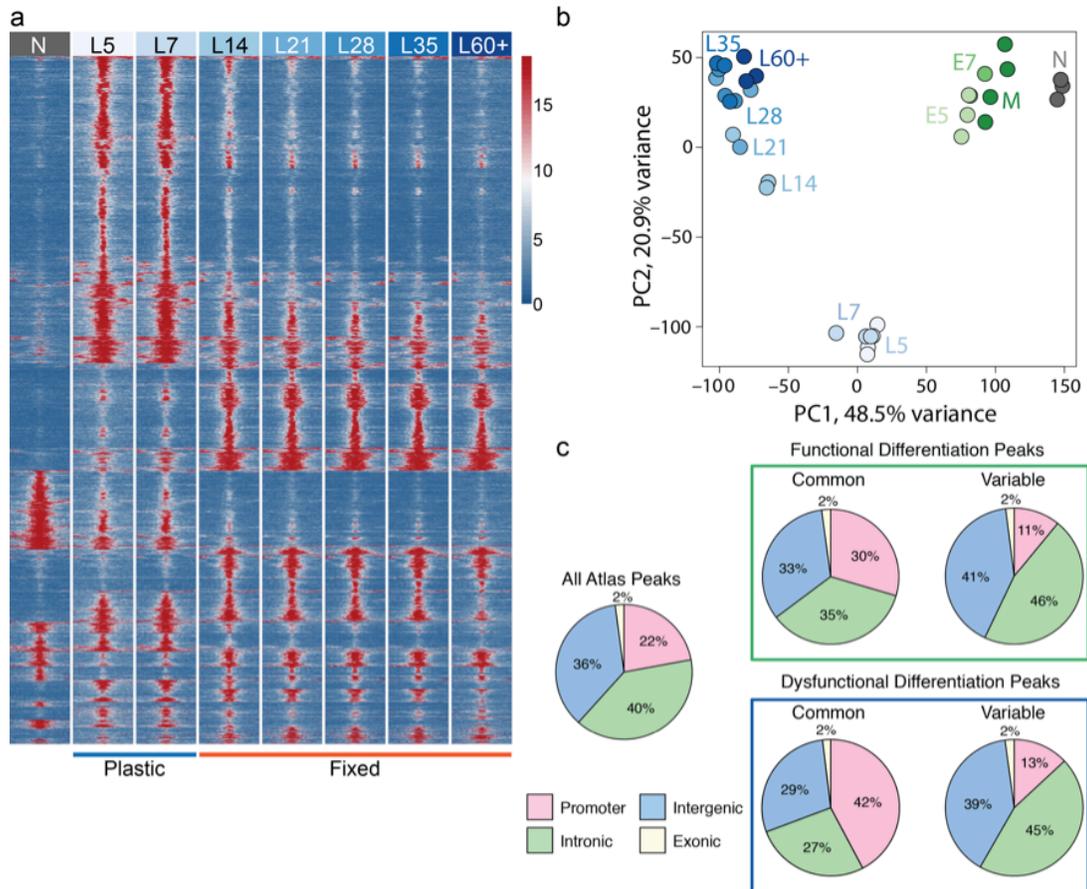


Figure 3.6: Overview of chromatin accessibility in TST

(a) Chromatin accessibility heat map (15,275 differentially accessible regions, FDR<0.05, absolute log₂ fold change > 1). **(b)** Principal component analysis (PCA) of peak accessibility in naive TCR_{TAG} cells (N; grey) during normal differentiation (green) and during tumorigenesis (blue). **(c)** Pie charts show proportions of reproducible ATAC-seq peaks in exonic, intronic, intergenic, and promoter regions (left, distribution for all peaks in the atlas). Green box: normal CD8 T cell differentiation during *Lm*TAG immunization; Blue box: differentiation to dysfunction in progressing tumors. Variable: significant change in at least one cell type comparison. Common: no significant change in any cell type comparison.

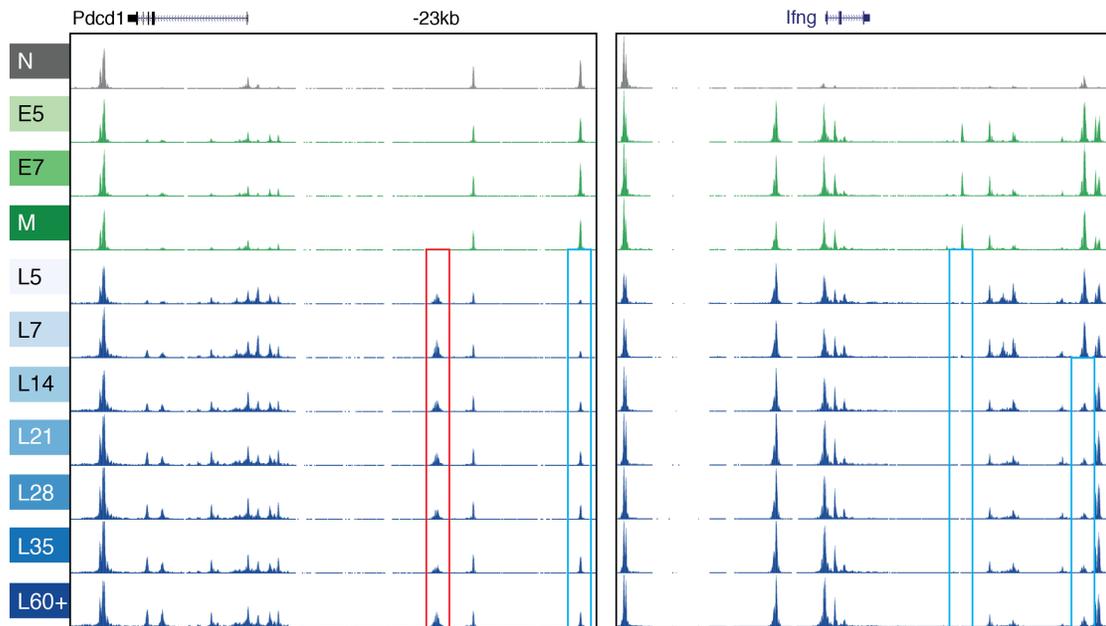


Figure 3.7: ATAC-seq signal in *Pdc1* and *Ifng* loci in normal and TST cells. Peaks present only in TST cells or normal cells are highlighted in red and blue, respectively.

To determine which transcription factors were associated with the transition from N to D5 relative to normal differentiation from N to E5, we tested whether accessibility of predicted transcription factor targets changed significantly in these transitions (**Figure 3.8a**). Predicted NFATC1-binding sites, including those associated with genes encoding inhibitory receptors and negative regulators such as *Ctla4*, *Pdc1*, *Tigit*, *Socs1*, and *Cblb* and transcription factors *Egr1* and *Egr2*, had increased accessibility in dysfunctional L5 relative to naïve cells (**Figure 3.8a**). NFAT transcription factor family members, particularly NFATC1 and NFATC2, are important regulators of T cell development and function⁷⁶, as well as exhaustion in chronic viral infections⁷⁷. Potential target genes whose binding sites opened in L5 relative to E5 (**Figure 3.8b**) were significantly more likely to be more highly expressed in L5 relative to E5 and the opposite was true for target genes containing binding sites that closed in L5 relative to E5 ($P < 1e-16$ for both comparisons, Fisher's exact test), consistent with

the hypothesis that differential accessibility of these NFATC1 binding sites leads to downstream differential expression of predicted targets.

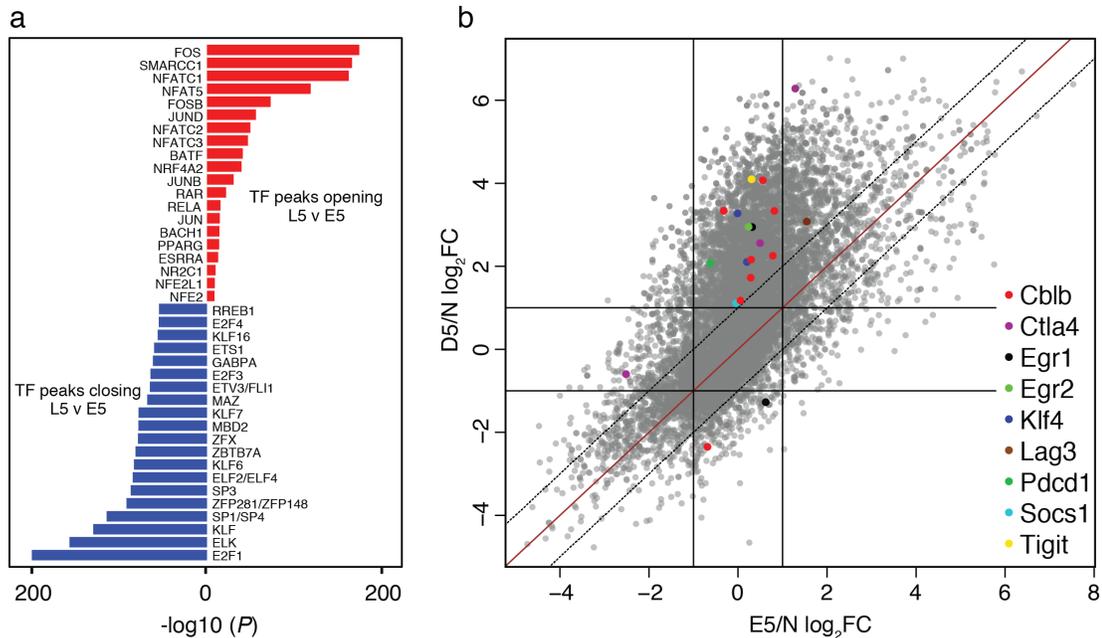


Figure 3.8: Transcription factor accessibility in naïve to L5 transition

(a) The 20 most significantly enriched transcription factor motifs in peaks opening (red) and closing (blue) between L5 and E5. **(b)** Scatterplot comparing the changes in peak accessibility for all peaks containing the NFATC1 motif during the transition from N to E5 TCR_{TAG} cells during acute listeria *LmTAG* infection versus N to L5 in pre-malignant lesions. Highlighted are NFATC1 target peaks associated with genes encoding negative regulatory transcription factors and inhibitory receptors. Some genes, for example, Cblb and Klf4, had multiple NFATC1 target peaks, including peaks that decreased in accessibility.

3.4 Chromatin states correlate with reprogrammability

Notably, the discrete chromatin states in dysfunctional TCR_{TAG} cells correlate temporally with our previous observation that L8 but not L35 were capable of regaining effector function³⁰. Indeed, when we re-isolated TCR_{TAG} cells from liver lesions and cultured them in vitro with IL-15 (**Figure 3.9a**), previously shown to induce

proliferation and restore effector function in tumor-reactive CD8 T cells^{78,79}, L5 and L7 regained the ability to produce IFN γ and TNF α , but TCR_{TAG} cells isolated at day 12 and after did not (**Figure 3.9a**). Thus the first dysfunctional state is plastic, but with concomitant chromatin remodeling between days 7 and 14, becomes fixed.

Genomic regions containing binding motifs of TCF family members closed during the transition from the plastic (L7) to fixed (L14) states, whereas sites containing motifs for E2F, ETS, and KLF family members opened (**Figure 3.9b**). Additionally, TCF1 (encoded by *Tcf7*) protein levels decreased between L7 and L14 (**Figure 3.9c**), and analysis of closing peaks showed enrichment for associated WNT receptor signaling pathway genes, regions upstream of TCF family transcription factors, as well as

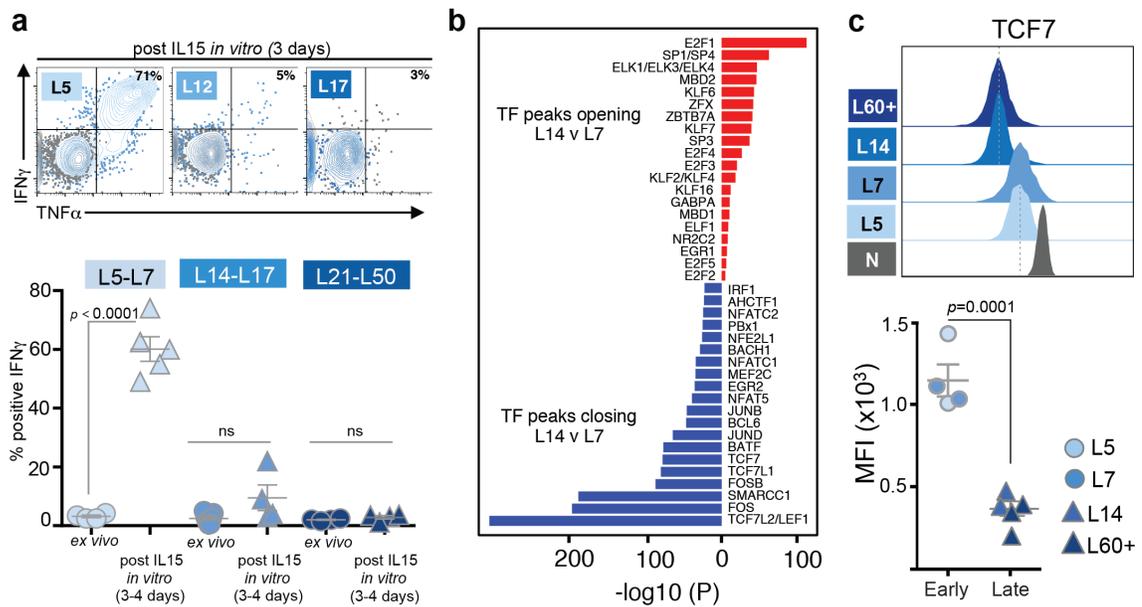


Figure 3.9: Tumor-associated T cell dysfunction occurs in two states

(a) Top, cytokine production by L5, L12, and L17 after 3 days *in vitro* IL-15 culture (grey, no peptide control). Bottom, IFN γ production *ex vivo* (circles) or after 3–4 days IL-15 *in vitro* culture (triangles). Pooled from three experiments. (b) The 20 most significantly enriched transcription factor motifs in peaks opening (red) and closing (blue) between L7 and L14. (c) TCF1 expression (MFI; mean fluorescence intensity). Each symbol represents individual mouse. Mean \pm s.e.m. shown

cytokine response, TCR signaling, and T cell differentiation pathway genes (**Table 3.1**). Among the TCR signaling genes most up-regulated during the L7–L14 transition were negative regulators such as *Cish1* and *Socs2*, whereas co-stimulatory molecule genes such as *Icos* and *Cd28* were down-regulated.

We next used an *in vivo* pharmacologic strategy to validate the predicted role of NFAT and TCF in TST dysfunction. FK506 is an immunosuppressant that inhibits NFAT nuclear translocation and downstream gene activation^{80,81}, and we used 25% of the full immunosuppression dose to partially down-regulate NFAT activity without completely blocking T cell activation and/or effector function. TWS119, a GSK3 β inhibitor, enhances differentiation of CD8 T cells to memory cells through WNT/TCF1 activation⁸². We treated TCR_{TAG}-adoptively transferred AST-Cre-ER^{T2} mice with FK506 alone or in combination with TWS119 (**Figure 3.10a**) and found that L10 TST from FK506 and FK506/TWS119-treated mice had decreased expression of the NFATC1 targets PD1 and LAG3, increased levels of TCF1 and EOMES (**Figure 3.10b**), and were more efficiently reprogrammable (**Figure 3.10c**) compared to controls or TWS119 alone (data not shown).

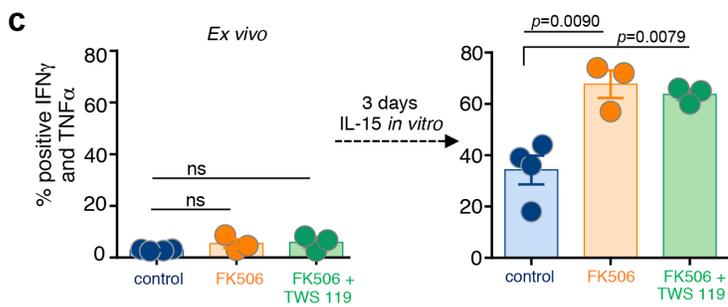
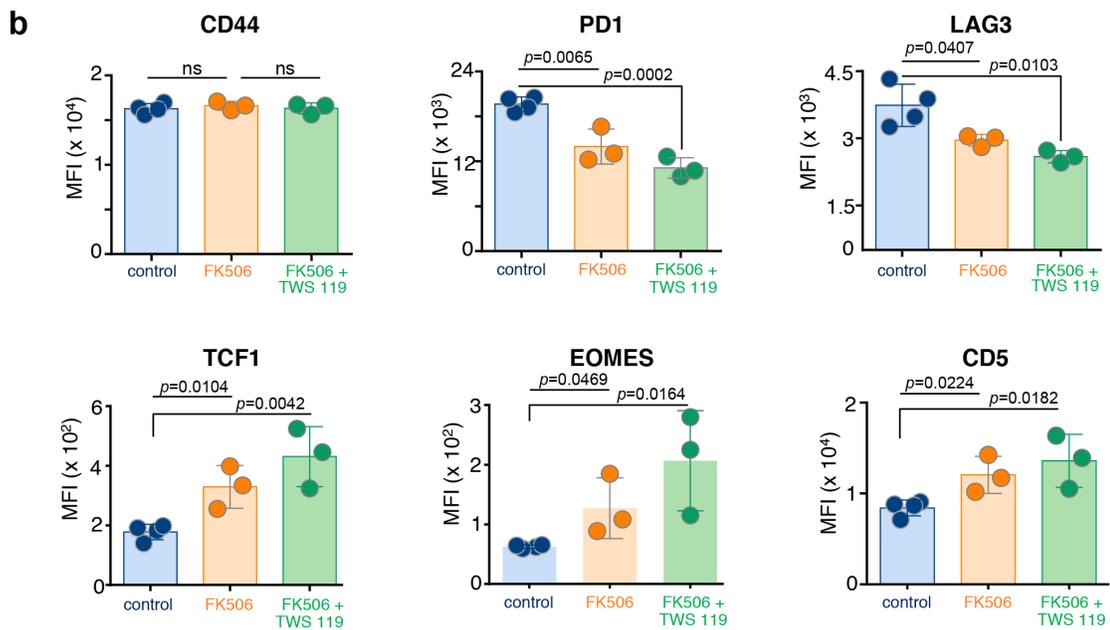
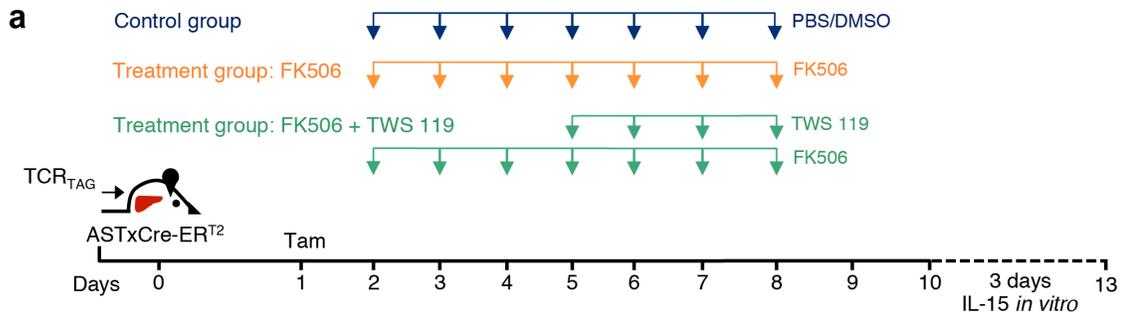
Table 3.1: GO terms enriched in regions closing in L7 to L14 transition

Selected enriched biological process gene ontology (GO) terms associated with chromatin regions which significantly lost chromatin accessibility during the L7 to L14 transition as determined by GREAT analysis.

Biological process GO term	Rank	FDR
Cellular response to cytokine stimulus	66	6.93E-14
Positive regulation of T cell activation	75	7.05E-13
Positive regulation of T cell differentiation	143	3.71E-10
Regulation of canonical Wnt receptor signaling pathway	231	4.77E-08
Positive regulation of T cell receptor signaling pathway	306	4.63E-07
Regulation of lymphocyte differentiation	312	5.38E-07
Regulation of interferon-gamma production	368	1.99E-06
Regulation of transcription regulatory region DNA binding	459	1.35E-05

Figure 3.10: Pharmacological targeting of NFAT and Wnt/ β -catenin signaling prevents TST differentiation to the fixed dysfunctional state *in vivo*.

(a) Experimental scheme. Naive TCR_{TAG} cells (Thy1.1+) were transferred into AST-Cre-ERT2 (Thy1.2+) mice which were treated with tamoxifen (tam) one day later. At days 2–9 mice were treated with the calcineurin inhibitor FK506 (2.5 mg kg⁻¹ per mouse) alone (FK506 treatment group; orange), or in combination with the GSK3 β inhibitor TWS119 (0.75 mg per mouse; days 5–8) (FK506 + TWS119 treatment group; green), or PBS/DMSO (control group; blue) as indicated. At day 10, TCR_{TAG} cells were isolated from livers and assessed for phenotype and function. (b) Flow cytometric analysis of CD44, PD1, LAG3, TCF1, and EOMES expression of TCR_{TAG} cells. (c) Production of IFN γ and TNF α by TCR_{TAG} cells isolated at day 10 (left panel; *ex vivo*), and after 3 days IL-15 *in vitro* culture (right panel). Each symbol represents an individual mouse. Data show mean \pm s.e.m.; *P* values calculated using unpaired two-tailed *t*-test.



3.5 Surface proteins associated with chromatin states

We wanted to determine if any surface proteins could be used to differentiate between TSTs in plastic and fixed dysfunctional states and thus indicate reprogrammability of heterogeneous TIL. PD1 and LAG3, two inhibitory receptors, are similarly expressed by both plastic (L5, L7) and fixed (L14+) dysfunctional TST (**Figure 3.5**) and thus not informative in this regard. We identified membrane protein genes differentially expressed between plastic (L5, L7) and fixed (L14 to L60+) dysfunctional TCR_{TAG} cells (**Figure 3.11a**) and found several markers not previously associated with tumor-induced T cell dysfunction. Plastic (L5, L7) TCR_{TAG} cells had low expression of CD38, CD101, and CD30L and high expression of CD5, whereas fixed (L14, L28) TCR_{TAG} cells had the opposite pattern (**Figure 3.11b**). Consistent with its expression, the *Cd38* locus contained intergenic and intronic ATAC-seq peaks only accessible in fixed-dysfunctional TST (**Figure 3.12**). Moreover, TCR_{TAG} cells from FK506 and FK506/TWS119-treated mice expressed low CD38 and CD101 compared to controls, correlating with their improved reprogrammability (**Figure 3.10d**). To test whether these markers could identify reprogrammable T cells within a heterogeneous TST population, we sorted CD38^{lo}CD101^{lo} and CD38^{hi}CD101^{hi} TST from PD1^{hi} L14 cells and assessed reprogrammability (3 days in vitro IL-15). CD38^{lo}CD101^{lo} L14 regained the ability to produce IFN γ and TNF α , but CD38^{hi}CD101^{hi} L14 did not (**Figure 3.11c**).

To determine if these findings could be applied to other tumor histologies and/or T cell specificities, we repeated these experiments using mouse B16F10 (B16) melanoma cells expressing ovalbumin (B16-OVA), a model antigen recognized by OVA-specific OT1 CD8 T cells (TCR_{OT1} cells). Naïve congenically marked TCR_{OT1} cells were adoptively transferred into B16-OVA tumor-bearing B6 mice. Tumor-infiltrating TCR_{OT1} cells up-regulated CD44, PD1, and LAG3, down-regulated CD62L,

and lost the ability to produce IFN γ and TNF α (**Figure 3.13a**). At later stages, dysfunctional TCR_{OT1} cells expressed high levels of CD38 and CD101 and down-regulated CD5 relative to day 5 dysfunctional TCR_{OT1} cells (**Figure 3.13b**). Moreover, late dysfunctional TCR_{OT1} cells at day 25 could not regain the ability to produce IFN γ or TNF α , in contrast to early dysfunctional TCR_{OT1} cells at day 5 (**Figure 3.13c**).

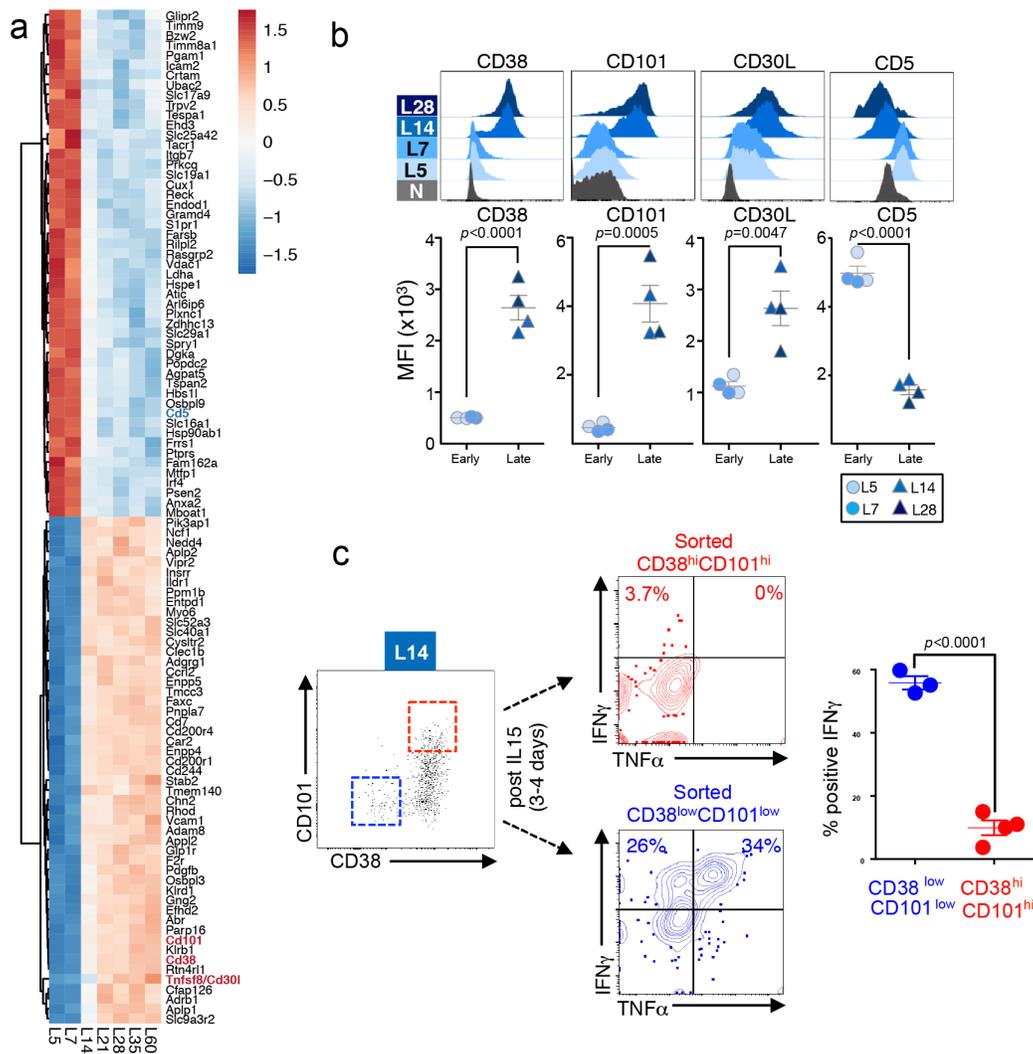


Figure 3.11: Cell surface markers for dysfunctional state transition

(a) RNA-seq expression (row-normalized) for the 50 most significant differentially expressed genes encoding membrane proteins. (b) CD38, CD101, CD30L and CD5 expression; representative of 3 independent experiments. (c) Cytokine production by sorted CD38^{lo}CD101^{lo} (blue) and CD38^{hi}CD101^{hi} (red) L14 after 3 days IL-15 *in vitro* culture. Similar data obtained with sorted L10 in independent experiment. Each symbol represents an individual mouse.

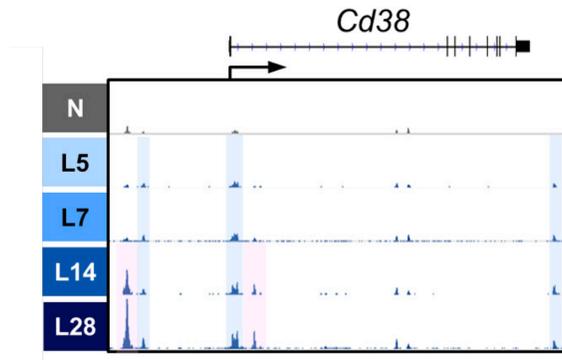


Figure 3.12: ATAC-seq signal profile across the *Cd38* locus.
Peaks accessible only in fixed dysfunctional cells highlighted in pink; activation-associated peaks highlighted in blue.

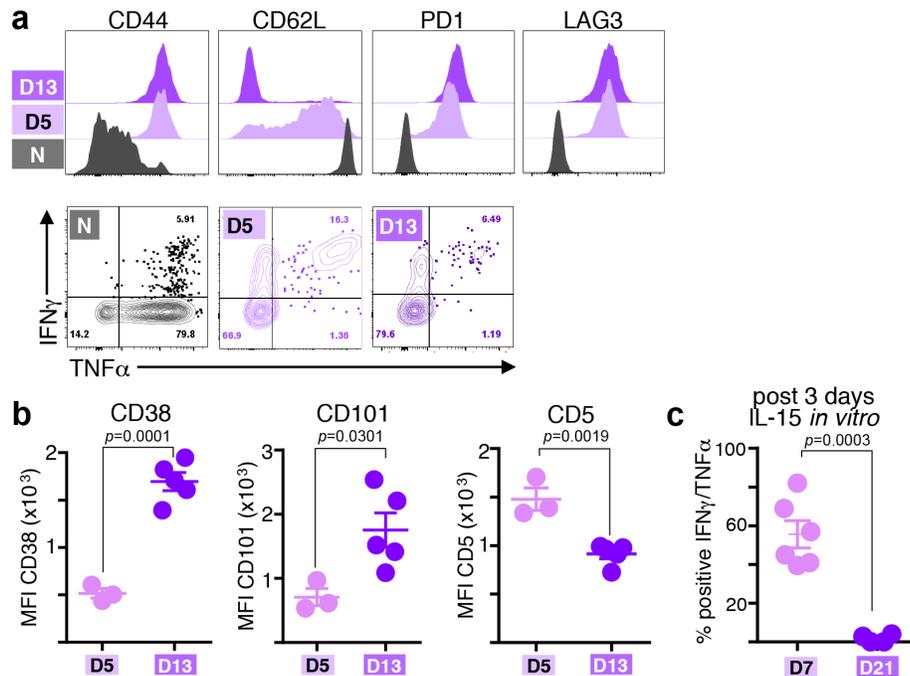


Figure 3.13: B16-OVA model

(a) Immunophenotype of and cytokine production by TCR_{OT1} cells re-isolated from established B16-OVA tumors 5 (D5) and 13 (D13) days after transfer. (b) CD38, CD101 and CD5 expression on day 5 and day 13 TCR_{OT1} cells. (c) Cytokine production by day 5 and day 21 TCR_{OT1} cells after 3 days of IL-15 *in vitro* culture. Each symbol represents an individual mouse. Mean ± s.e.m. shown.

3.6 Memory T cells enter fixed dysfunctional state in tumor

As memory cells (M; **Figure 1.4a**) have been previously activated and their chromatin accessibility landscape is similar to that of functional effector cells, we hypothesized that they may be resistant to the differentiation to dysfunction we observed in naïve cells in the tumor environment. We transferred TCR_{TAG} memory cells into AST-Alb-Cre^{T2} mice (in which hepatocytes express TAG from birth³⁰) bearing established hepatocellular carcinomas and, one day later, immunized with *LmTAG* (**Figure 3.14a**). By day 7, tumor-infiltrating memory T cells (ML7) rapidly up-regulated PD1 and LAG3 and progressively lost effector function (**Figure 3.14b**). ATAC-seq revealed that M cells followed a similar epigenetic trajectory as the N cells in early malignant lesions (**Figure 3.14c,d**) and remarkably, by day 35, the chromatin state of transferred M cells was nearly identical to that of N at day 35 in early malignant lesions (ML35 and L35, respectively; **Figure 3.14d**). Dysfunctional M cells displayed the same gain and loss of ATAC-seq peaks in critical gene loci including *Pdcd1*, *Ctla4*, *Cd38*, *Tcf7*, and *Ifng* (**Figure 3.15a**). Changes in surface protein expression (CD38, CD101, CD30L, and CD5) between ML7 and ML14 were similar to those seen with N (L7 and L14, respectively) (**Figure 3.15b**). We obtained similar results when *LmTAG* immunization after adoptive transfer was omitted (**Figure 3.14c,d**).

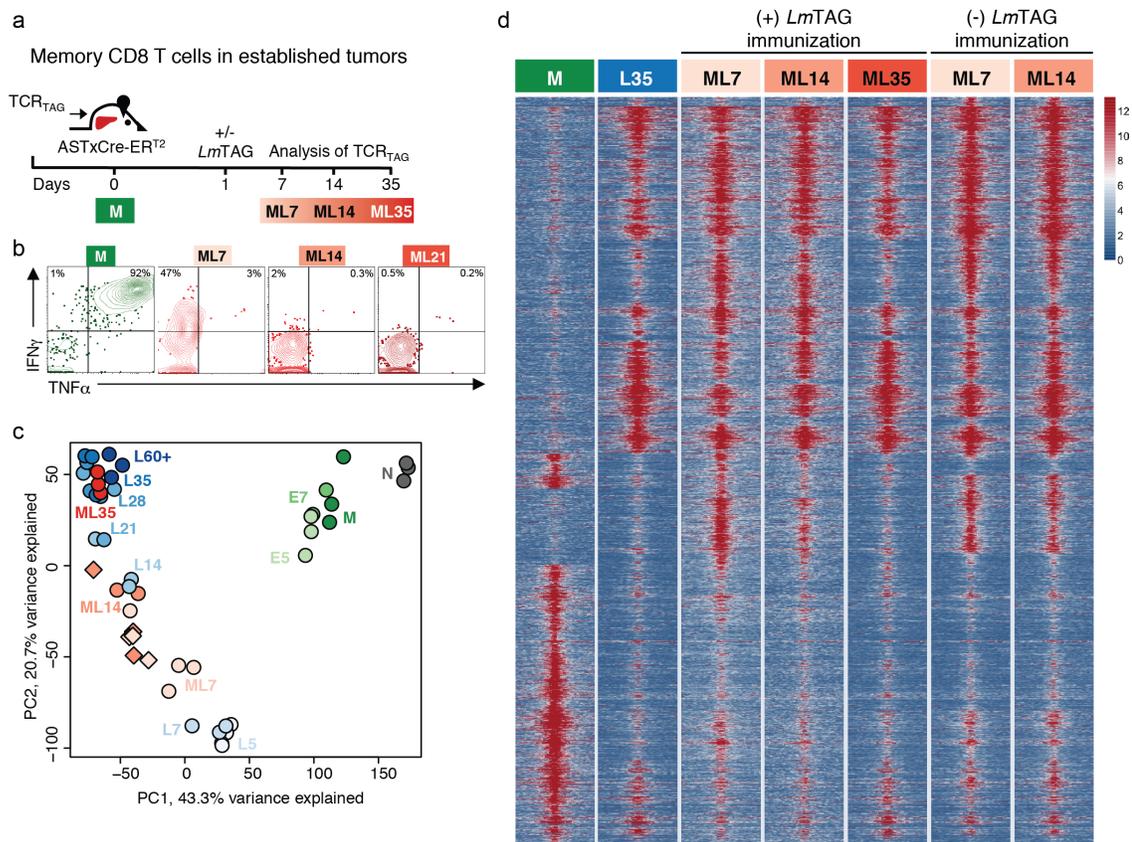


Figure 3.14: Memory CD8 T cells in established tumors

(a) Experimental scheme **(b)** Cytokine production of M cells isolated from liver tumors. **(c)** PCA of peak accessibility in TCR_{TAG} cells during acute infection (green), tumorigenesis (blue), and memory TCR_{TAG} cells in established tumors (red). Diamonds represent adoptively transferred memory cells without *LmTAG* immunization **(d)** Chromatin accessibility heat map. Each row represents 1 of 11,698 selected peaks (differentially accessible between any sequential cell comparison; FDR < 0.05, log₂(FC) > 2). Shown are \pm 1kb from the peak summit (2kb total per region).

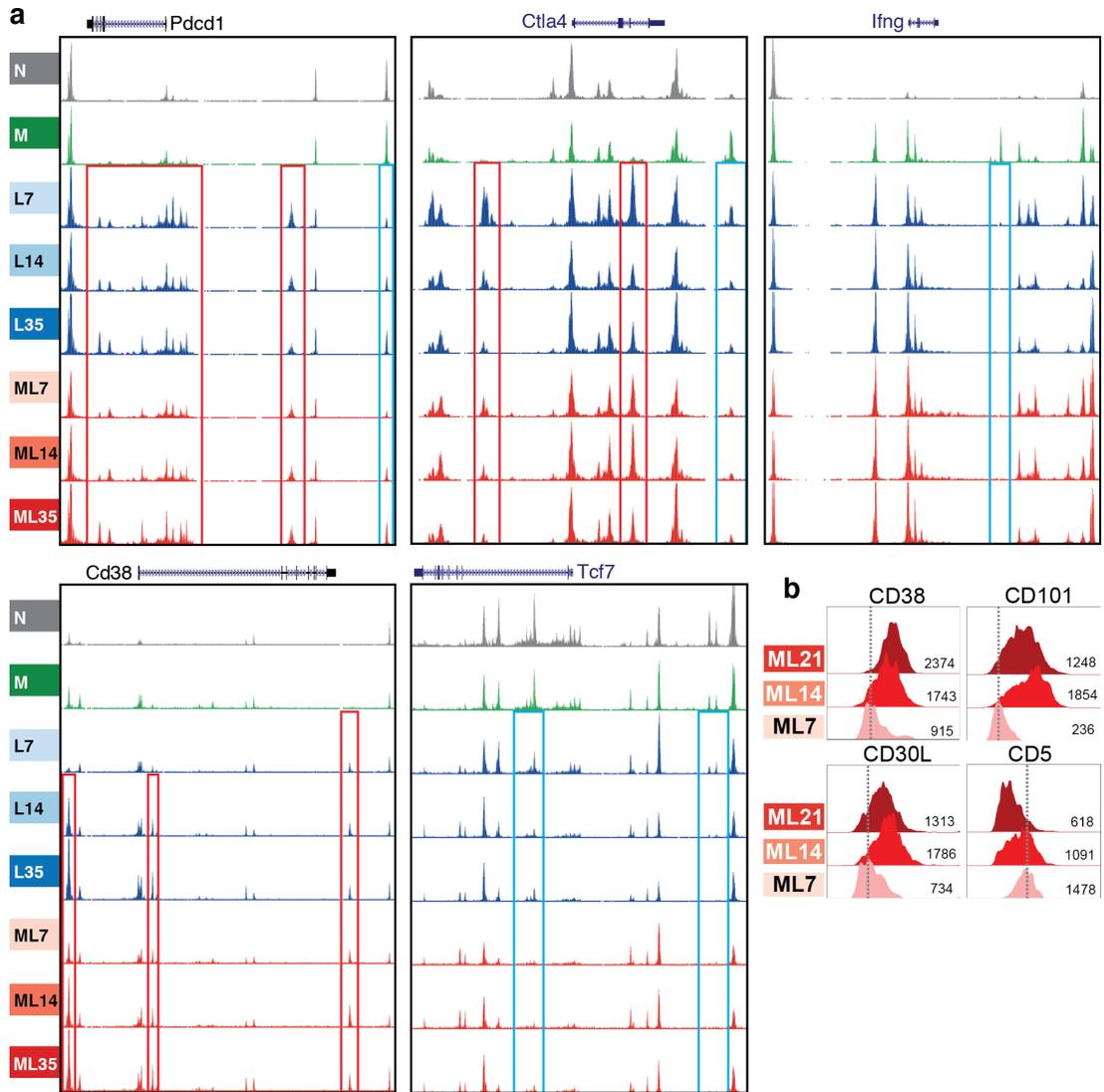


Figure 3.15: Dysfunctional memory cells

(a) ATAC-seq signal profiles of *Pdc1*, *Ctla4*, *Cd38*, *Tcf7*, and *Ifng* genes for naive (N; grey), memory (M; green), L7, L14, L35 (blue series), and ML7, ML14, and ML35 (red series) TCR_{TAG} cells. Red boxes highlight peaks that become accessible in dysfunctional T cells compared to naive and memory; blue boxes highlight peaks that become inaccessible in dysfunctional TCR_{TAG} cells compared to naive and memory.

(b) CD38, CD101, CD30L, and CD5 expression in ML7, ML14, ML21. Inset numbers show MFI.

3.7 Chromatin accessibility in human TILs

Finally, we examined chromatin states of human CD8 TIL and peripheral blood lymphocytes from healthy donors. We carried out ATAC-seq on naive (N; CD45RA⁺CD45RO⁻), effector memory (EM; CD45RA⁻CD45RO⁺CD62L^{lo}), and central memory (CM; CD45RA⁻CD45RO⁺CD62L^{hi}) CD8 peripheral blood lymphocytes from healthy donors and PD1^{hi} CD8 TIL isolated from human melanoma and non-small-cell lung cancer tumors (**Figure 3.16a**). Human N cells had a distinct chromatin state as compared to EM and CM, which were similar (**Figure 3.16b,c**),

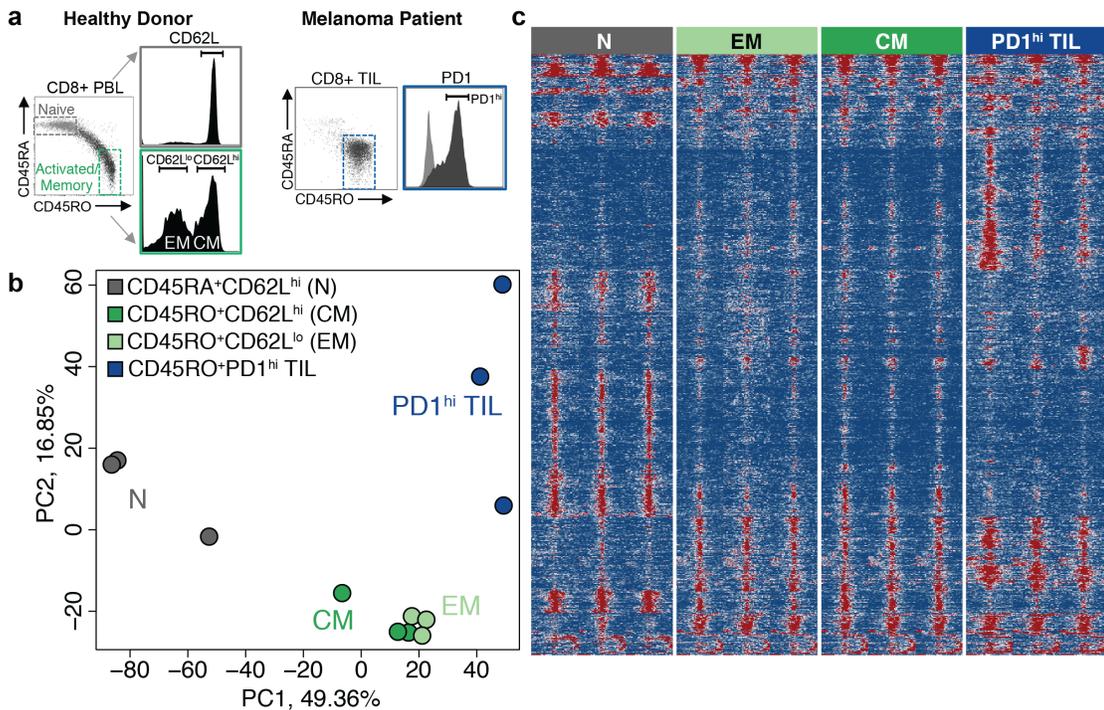


Figure 3.16: Normal human T cells and tumor-infiltrating lymphocytes
(a) Peripheral blood lymphocyte sorting scheme for naive (N), effector memory (EM), central memory (CM) CD8 T cell populations (left), and PD1^{hi} CD8 TIL from patients with melanoma or non-small-cell lung cancer. **(b)** PCA using normalized read counts from all reproducible ATAC-seq peaks in human healthy donor peripheral blood lymphocytes and PD1^{hi} TIL from melanoma and non-small-cell lung cancer tumors. **(c)** Differentially accessible ATAC-seq peaks grouped by DESeq-defined differential accessibility pattern. Each column represents one biological replicate.

though distinct accessibility patterns in genes such as *SELL*. Multiple peaks were gained or lost only in PD1^{hi} TIL, including peaks in *IFNG*, *EGR2*, *CD5*, and *CTLA4* (**Figure 3.17a**). We compared the accessibility changes in peaks outside of promoter regions that occurred during functional and dysfunctional mouse CD8 T cell differentiation with those observed in human peripheral blood lymphocytes and PD1^{hi} TIL and found that human PD1^{hi} TIL had the greatest correlation in peak accessibility changes with fixed dysfunctional (late-stage) mouse TST, relative to their respective naïve cells (**Figure 3.17b**). For example, the *TCF7/Tcf7* locus showed similar intergenic and intronic peak accessibility changes in human PD1^{hi} TIL and mouse fixed-dysfunctional TCR_{TAG} cells (**Figure 3.17c**). A subset of PD1^{hi} TIL expressed higher levels of CD38 and CD101 and lower levels of CD5 (**Figure 3.18**), suggesting that these markers could potentially be used to identify T cells that are amenable to therapeutic reprogramming in human tumors.

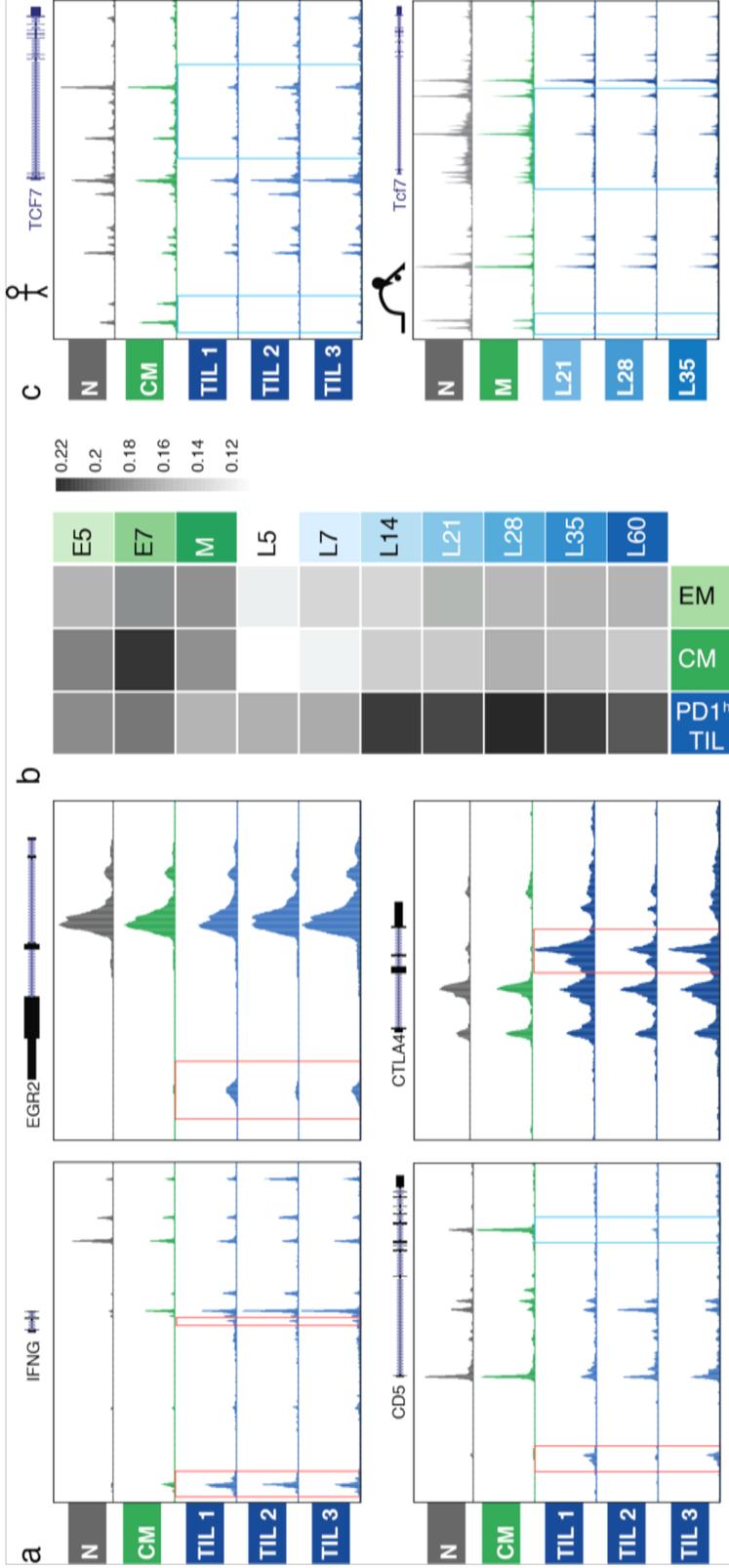


Figure 3.17: Comparison of mouse and human chromatin accessibility

(a) ATAC-seq signal profiles of *IFNG*, *EGR2*, *CD5*, and *CTLA4*. Red and blue boxes highlight peaks that become accessible or inaccessible, respectively, in PD1^{hi} TIL (blue) compared to naïve (grey) or central memory (green). (b) Heatmap of Spearman correlations of non-promoter peak accessibility fold changes between human N and EM, CM or PD1^{hi} TIL and accessibility fold changes between mouse N and E5, E7, M, and L5 to L60. $P < 10^{-16}$ for all comparisons between human PD1^{hi} TIL and mouse L14–L60. (c) ATAC-seq signal profiles across human *TCF7* and mouse *Tcf7* gene loci; peaks lost in human PD1^{hi} TIL and mouse L21, L28, L35 highlighted in blue.

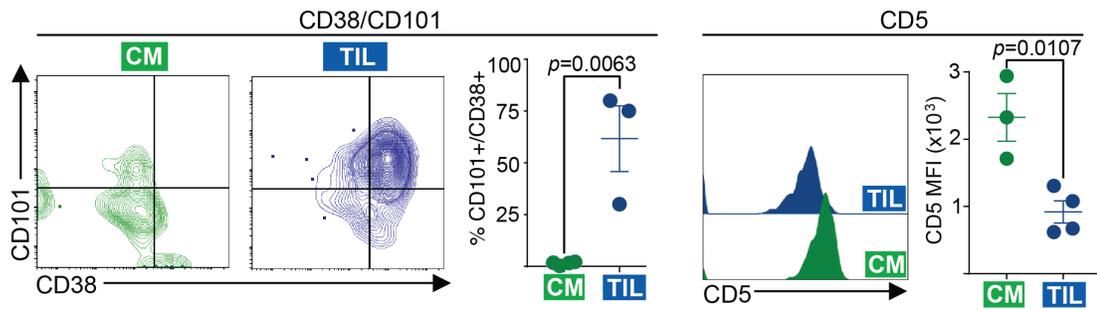


Figure 3.18: Characterization of normal human T cells and TILs. CD38, CD101 and CD5 expression in human CM (green) and PD1^{hi} TIL (blue). Each symbol represents an individual healthy donor/patient. Mean \pm s.e.m. shown.

3.8 Discussion

In this study, we define the chromatin state dynamics underlying tumor-specific T cell dysfunction over the course of tumorigenesis. Naive TST encountering tumor antigen in pre-malignant lesions differentiated to an initially plastic, therapeutically reprogrammable chromatin state, then transitioned to a fixed dysfunctional chromatin state that did not undergo further remodeling, even after progression to large established tumors. The rapid induction of dysfunction early during tumorigenesis without progression through an effector state resembles peripheral self-tolerance induction^{83,84}. Surprisingly, memory TST differentiated to the same fixed dysfunctional chromatin state in tumors, suggesting that antigen exposure in tumors can overwrite pre-existing epigenetic programs regardless of the initial differentiation state.

We identified surface markers, including CD101 and CD38, which were associated with discrete dysfunctional chromatin states and demarcated reprogrammable from non-reprogrammable PD1^{hi} T cells within heterogeneous TIL populations, a finding of important potential clinical relevance, and human PD1^{hi} TIL showed heterogeneous expression of these markers. In patients who do not respond to immune checkpoint blockade (non-responders), PD1^{hi} TIL may be in a fixed dysfunctional state,

in contrast to responders whose PD1^{hi} TIL are in a plastic state, amenable to reprogramming. Our studies on the epigenetic and transcriptional programs underlying TST dysfunctional states and therapeutic reprogrammability point to new targets and strategies to transform TST into potent anti-tumor agents.

3.9 Methods

ATAC-seq processing pipeline: Raw ATAC-seq reads were trimmed and filtered for quality using Trim Galore! v0.4.0⁸⁵, powered by CutAdapt v1.8.1⁸⁶ and FastQC v0.11.3⁸⁷. Paired-end reads were aligned using Bowtie2 v2.2.5⁸⁸ against either mm10 or hg38 and non-uniquely mapping reads were removed. To correct for the fact that the Tn5 transposase binds as a dimer and inserts two adapters in the Tn5 tagmentation step⁸⁹, all positive-strand reads were shifted 4bp downstream and all negative-strand reads were shifted 5bp upstream to center the reads on the transposon binding event³². We then pooled the shifted reads by sample type and identified peaks using MACS2⁹⁰ with a threshold of FDR-corrected $P < 1 \times 10^{-2}$ using the Benjamini–Hochberg procedure for multiple hypothesis correction. As called peaks may be caused by noise in the assay and not reflect true chromatin accessibility, we calculated an irreproducible discovery rate (IDR)⁹¹ for all pairs of replicates across a cell type. The IDR is an estimate of the threshold where two ranked lists of results, in this case peak calls ranked by P value, no longer represent reproducible events. Using this measure, we excluded peaks that were not reproducible ($IDR < 5 \times 10^{-3}$) across at least one pair of replicates in each mouse or human cell type.

ATAC-seq atlas creation: Peaks found reproducibly in each mouse cell type were combined to create a genome-wide atlas of accessible chromatin regions. Reproducible

peaks from different samples were merged if they overlapped by more than 75%. To create the atlas of accessible peaks for the human samples, reproducible peaks from the normal human cell types (HN, HCM, and HEM) and the tumor-derived cells (PD1^{hi}) were combined. There was greater variation between the human TIL samples than between T cell samples from healthy donors; this led to fewer reproducible peaks being called in the TIL samples. Like the mouse atlas, peaks overlapping by more than 75% were merged in the human atlas. Numbers of called peaks and reproducible peaks for each sample type are listed in **Table 3.2**.

Assignment of ATAC-seq peaks to genes: The RefSeq transcript annotations of the hg38 version of the human genome and the mm10 version of the mouse genome were used to define the genomic location of transcription units. For genes with multiple gene models, the longest transcription unit was used for the gene locus definition. ATAC peaks located in the body of the transcription unit, together with the 2-kb regions upstream of the TSS and downstream of the 3' end, were assigned to the gene. If a peak was found in the overlap of the transcription units of two genes, one of the genes was chosen arbitrarily. Intergenic peaks were assigned to the gene with a TSS or 3' end that was closest to the peak. In this way, each peak was unambiguously assigned to one gene. Peaks were annotated as promoter peaks if they were within 2kb of a transcription start site. Non-promoter peaks were annotated as intergenic, intronic or exonic according to the relevant RefSeq transcript annotation.

Table 3.2: Number of peaks per cell type

Number of called ATAC-seq peaks (MACS2 $p=0.01$) and number of reproducible peaks (IDR<0.001) for each cell state.

	Called peaks	Reproducible peaks
N	91,255	46,089
E5	87,914	44,520
E7	109,081	49,625
M	100,110	27,191
L5	105,022	43,960
L7	92,692	41,322
L14	91,396	44,648
L21	79,556	40,264
L28	84,994	45,055
L35	82,536	44,660
L60+	76,228	38,316
ML7	110,847	38,753
ML14	86,946	30,798
ML35	81,446	44,508
ML7_noLM	101,478	45,844
ML14_noLM	95,763	51,534
Human N	133,716	32,675
Human CM	91,745	36,283
Human EM	88,521	29,666
Human PD1hi_1 TIL	73,748	-
Human PD1hi_2 TIL	87,029	-
Human PD1hi_3 TIL	98,554	-

ATAC-seq peak atlas summary: We found a total of 75,689 reproducible ATAC-seq peaks in the mouse samples. Examining genomic locations, 39.6% of the peaks were found in introns, 36.3% were found in intergenic regions, 22.1% were found in promoters and 2.1% were found in exons. In the human samples, we found a total of 42,104 reproducible ATAC-seq peaks. Among these peaks, 34.0% were found in introns, 29.9% were found in intergenic regions, 34.0% were found in promoters, and 2.0% were found in exons. Chromosome-wide genomic coverage for all (autosomal) chromosomes and all samples was examined and no systemic bias was observed.

Principal component analysis: PCA plots were generated using read counts against all mouse or human atlas peaks. These read counts were processed using the variance-stabilizing transformation built into the DESeq2 package⁹².

Differential peak accessibility: Reads aligning to atlas peak regions were counted using the summarizeOverlaps function of the R packages GenomicAlignments v1.2.2 and GenomicRanges v1.18.4⁹³. Differential accessibility of these peaks was then calculated for all pairwise comparisons of cell types using DESeq2 v1.6.3⁹².

Peak heat maps and genome coverage plots: The ATAC-seq peak heat maps were created by pooling the DESeq size-factor normalized read counts per atlas peak across replicates of ATAC-seq data and binning the region ± 1 kb around the peak summit in 20bp bins. To improve visibility, bins with read counts greater than the 75th percentile+1.5 \times IQR were capped at that value. All analysis was performed using the original uncapped read counts. Genome coverage plots were generated for each replicate of ATAC-seq and RNA-seq by calculating genome-wide coverage of aligned reads using the bedtools function genomcov⁹⁴. For ATAC-seq samples, this coverage

was calculated after shifting the reads to account for the Tn5-induced bias. The coverage values were then normalized using DESeq2-derived size factors and replicates were combined to create one signal track for each sample type. ATAC-seq and RNA-seq coverage plots were generated using the Integrated Genomics Viewer⁹⁵.

Transcription factor peak assignment: Using the MEME⁹⁶-curated CisBP⁹⁷ transcription factor binding motif (TFBM) reference, we scanned the mouse ATAC-seq peak atlas with FIMO⁹⁸ to find peaks likely to contain each TFBM ($P < 10^{-4}$). The MEME cisBP reference for direct and inferred motifs for *Mus musculus* was curated by the MEME suite developers as follows: to reduce redundancy, for each transcription factor a single motif was selected according to the following precedence rules: The direct motif was chosen if there was one, otherwise the inferred motif with the highest DNA binding domain (DBD) similarity (according to CisBP) to a transcription factor in another species with a direct motif was chosen. If there was more than one direct motif or inferred motif with the highest DBD similarity, a motif was chosen according to its provenance (CisBP 'Motif_Type' attribute) in the following order: ChIP-seq, HocoMoco, DeBoer11, PBM, SELEX, B1H, High-throughput Selex CAGE, PBM:CSA:DIP-chip, ChIP-chip, COMPILED, DNaseI footprinting. Each motif thus determined was linked to a single transcription factor in the CisBP database, following the same precedence rules. The final reference contained 718 motifs between 6 and 30 bp in width (average width, 10.7 bp). Transcription factors with similar FIMO-predicted target peaks were combined into transcription factor families. Similarity of predicted target peak sets was measured using the Jaccard index (size of intersection/size of union). Transcription factors with Jaccard indices greater than 0.7 were combined for further analyses. Relative transcription factor accessibility was calculated using two one-sided Wilcoxon rank-sign tests comparing the distributions of

peak heights for peaks containing FIMO-predicted transcription factor binding sites. Peak height was defined as the maximum observed number of reads overlapping at any point in the defined peak region.

Comparison of human and mouse ATAC-seq atlases: The UCSC liftOver tool⁹⁹ was used to convert the mouse ATAC-seq peak atlas from mm10 coordinates to hg38 coordinates. The converted mouse atlas was then compared to the human atlas and 20,642 mouse peaks were within 100 bp of a human peak. We compared the results from the UCSC liftOver tool and an alternative method, bnMapper¹⁰⁰, and confirmed that the set of peaks mapped by bnMapper and by the UCSC liftOver tool was nearly identical (57,383 out of 75,689 by liftOver and 58,299 out of 75,689 by bnMapper). Additionally, all 57,223 peaks mapped to hg38 by both tools were mapped to the same chromosomal positions. The majority of these conserved peaks were found in promoter regions (56.4%), whereas relatively fewer were found in intergenic (22.4%), intronic (19.6%), and exonic (1.5%) regions. For non-promoter peaks conserved between human and mouse, Spearman correlations of $\log_2(\text{FC})$ were calculated between human N and human EM, CM or PD1^{hi} TIL versus $\log_2(\text{FC})$ between mouse N and functional E5, E7, M and dysfunctional L5 to L60.

RNA-seq: Raw RNA-seq reads were trimmed and filtered for quality using Trim Galore! v0.4.0⁸⁵, powered by CutAdapt v1.8.1⁸⁶ and FastQC v0.11.3⁸⁷. Paired-end reads were aligned using STAR¹⁰¹ against either mm10 or hg38. The RefSeq transcript annotations of the hg38 version of the human genome and the mm10 version of the mouse genome were used for the genomic location of transcription units. Reads aligning to annotated exon regions were counted using the summarizeOverlaps function of the R packages GenomicAlignments v1.2.2 and GenomicRanges v1.18.4⁹³.

Differential expression of genes across cell types was calculated using DESeq2⁹² v1.6.3. FDR correction of 0.05 was imposed unless otherwise stated. A log₂ fold change cutoff of 1 was used in some analyses as indicated.

Pathway analysis: Enrichment of gene ontology terms in sets of ATAC-seq peaks was calculated using GREAT (Genomic Regions Enrichment of Annotations Tool) using default parameters¹⁰². The full ATAC-seq atlas was used as the background set.

Membrane protein analysis: To identify membrane proteins that distinguished early (L5–L7) from late (L14–L60) dysfunctional TST, RNA-seq data was analyzed for genes contained within the gene ontology¹⁰³ category 0016020 (membrane proteins). The top 50 most up- and down-regulated genes (size-factor normalized RPKM) when compared between L5–L7 and L14–L60 were plotted in a heat map (row-normalized). Protein expression was assessed by flow cytometry for those membrane proteins for which monoclonal antibodies were available. Mouse targets (clone; supplier): CD5 (53-7.3; eBioscience), CD30L (RM153; eBioscience), CD38 (90; Biolegend), and CD101 (Moushi101; eBioscience). Human targets: CD5 (L17F12; Biolegend), CD38 (HB7; eBioscience), CD101 (BB27; Biolegend)

Data availability: All data generated and supporting the findings of this study are available within the paper. The RNA-seq and ATAC-seq data have been deposited in the Gene Expression Omnibus (GEO Super-Series accession number GSE89309 (GSE89307 for RNA-seq, GSE89308 for ATAC-seq).

CHAPTER 4

MODELING OF TRANSCRIPTION FACTOR ACTIVITIES TO PREDICT CHANGES IN GENE EXPRESSION

4.1 Introduction

As discussed previously, prolonged exposure of a T cell to its antigen can induce a dysfunctional state wherein the cell no longer produces common inhibitory molecules and the transcriptional state of the cell changes²⁹. These cells are no longer able to induce apoptosis in their target cells. This state can be induced by both cancer, where the antigen is a neoantigen, or by a chronic viral infection, where the antigen is a viral protein. However, although both of these antigens may induce a dysfunctional state, the antigens themselves arise through different mechanisms and it is not clear if these differences influence the initiation and progression of this dysfunctional program. Additionally, although the dysfunction induced by these two types of antigens appears similar, we do not know if these processes are the same below the surface, for example, at the epigenetic level. In this chapter I will describe my work attempting to address these questions.

4.2 T cell maturation and self-tolerance mechanisms

As it is critical that T cells do not activate aberrantly, there are complex central and peripheral mechanisms in place to ensure that mature T cells tolerate self-proteins¹⁰⁴. When T cells are being developed in the thymus, they are exposed to self-proteins loaded into the major histocompatibility complex (MHC). T cells that are strongly reactive to these self-proteins are deleted through apoptosis or are modified in other

ways. T cells that are not reactive at all are ignored and do not continue developing. Only the T cells that are slightly reactive to the self-proteins are able to mature, ensuring that functional, but not self-reactive, T cells are produced.

As not *all* self-antigens are expressed in the thymus, mechanisms of central tolerance are not completely effective. Additional protective mechanisms exist outside the thymus to block the activity of self-reactive T cells that did not encounter their self-antigen during development. In secondary immune organs such as the lymph nodes and spleen, immature dendritic cells (tolerogenic DCs) present additional self-antigens to T cells without a co-stimulatory signal. The lack of a stimulatory signal forces the T cell into a long-term inactive state, anergy. Alternatively, DCs expressing BTLA can induce the self-reactive T cell to differentiate to a regulatory T cell (Treg)¹⁰⁵.

Both cancer and chronic viral infections can induce a dysfunctional state similar to anergy. What is not clear however, is how similar these states are, if their initiations and progressions are similar, and if the resulting dysfunction is reversible.

3.3 Computational modeling of two dysfunctional T cell states

In order to compare the nonfunctional states induced in tumor-specific T cells (TSTs) and T cells in a chronic infection environment, we compared their epigenetic and transcriptional states. We modeled the activities of expressed transcription factors and determined their relative importance during various cell state transitions. To model the TSTs, we used the RNA-seq and ATAC-seq data described in Chapter 3 for normal CD8⁺ cells and for the cells in the tumor environment. To model the viral infection, we used RNA-seq and ATAC-seq data for CD8⁺ T cells from the mouse model for chronic lymphocytic choriomeningitis virus (LCMV) infection from a study performed in the Wherry lab¹⁰⁶. In this study, both RNA-seq and expression microarrays were used to

characterize gene expression changes. In order to minimize differences between the tumor and viral samples, we only used the viral samples with matching RNA-seq data (**Table 4.1**).

Table 4.1: Samples used in modeling

RNA-seq and ATAC-seq samples used in modeling for tumor and chronic viral infection environments

	Tumor	Viral
Naïve	X	X
Effector	X	
Memory	X	
Dysfunctional / Exhausted	X	X

A unified atlas for LCMV and TST cells was generated in order to minimize experiment and cell-type specific differences in chromatin accessibility across the samples. This atlas was created by first calling reproducible peaks for each cell type and experiment separately and then combining the peaks across experiments. The accessibility measures for the peaks were then batch-corrected by training a generalized linear model (GLM) on the data and including the batch as a confounding factor. The portion of the accessibility value that could be attributed to the batch could then be subtracted out, leaving values that could more accurately be compared across experiments. **Figure 4.1** shows a PCA of the combined data and it appears from this analysis that the normal naïve and effector cells are similar across experiments while the dysfunctional and exhausted cell types are also similar although their environments differ.

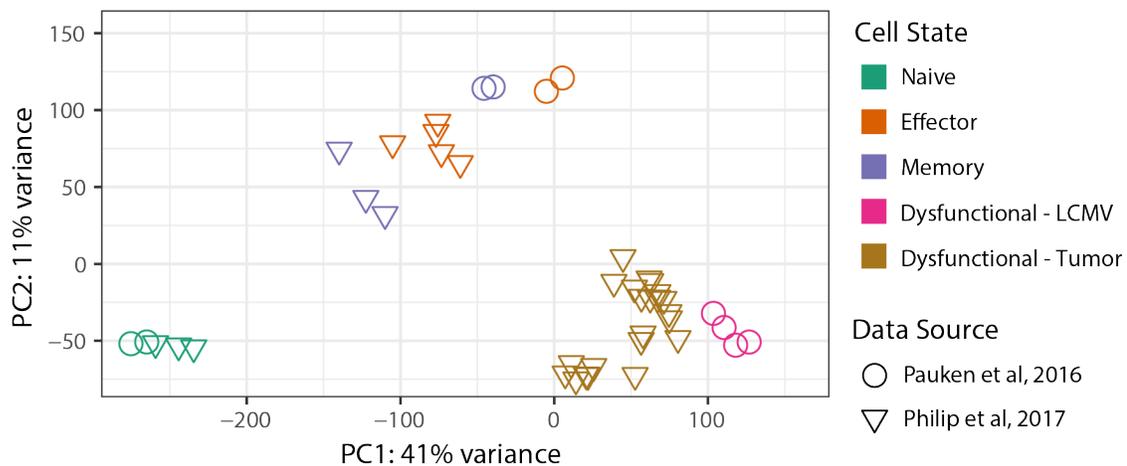


Figure 4.1: PCA of CD8+ T cell types.

First two components from a principal component analysis using the accessibility values of the top 10,000 most accessible peaks in any cell type after batch correction, comparing normal CD8+ T cells with T cells in LCMV and tumor environments

Similar to the work described in the previous chapter, we used cisBP’s defined sequence binding motifs⁹⁶ for transcription factors expressed in the relevant RNA-seq data to define potential transcription factor binding events within regions of accessible chromatin genome-wide. Genes were defined as “potentially-regulated” by a transcription factor if there was a cisBP-defined binding site for that TF within one or more chromatin accessibility peaks assigned to that gene. As the location of an accessible region of chromatin relative to a protein-coding gene is known to impact the type of regulation, atlas peaks were split into four groups for modeling purposes: “promoter peaks”, “opening peaks”, “closing peaks”, and “stable peaks”. Peaks were split based on their change in accessibility in the cell type transition being modeled, leading to the opening, closing, and stable groups. Promoter peaks were separated into

their own group since they are often constitutively accessible. Using this framework, a gene-by-TF matrix was created and used as input to a generalized linear model. The goal of this model was to predict gene expression change from ATAC-seq and TF binding data in order to learn the relative importance of each expressed TF in orchestrating that change.

This modeling approach was able to capture changes in RNA expression to varying degrees, depending on the cell type transition being modeled (**Figure 4.2**). The transition from dysfunctional day 7 to day 14 was one of the more accurately predicted transitions (Spearman correlation = 0.592, N=1076) whereas expression changes occurring during cell transitions with fewer chromatin changes, such as effector to memory cells, were not captured as well (Spearman correlation = 0.319, N=373). Across all transitions, genes with more reproducible open chromatin regions in any cell type (potentially more tightly regulated genes) were predicted more accurately (**Table 4.2**).

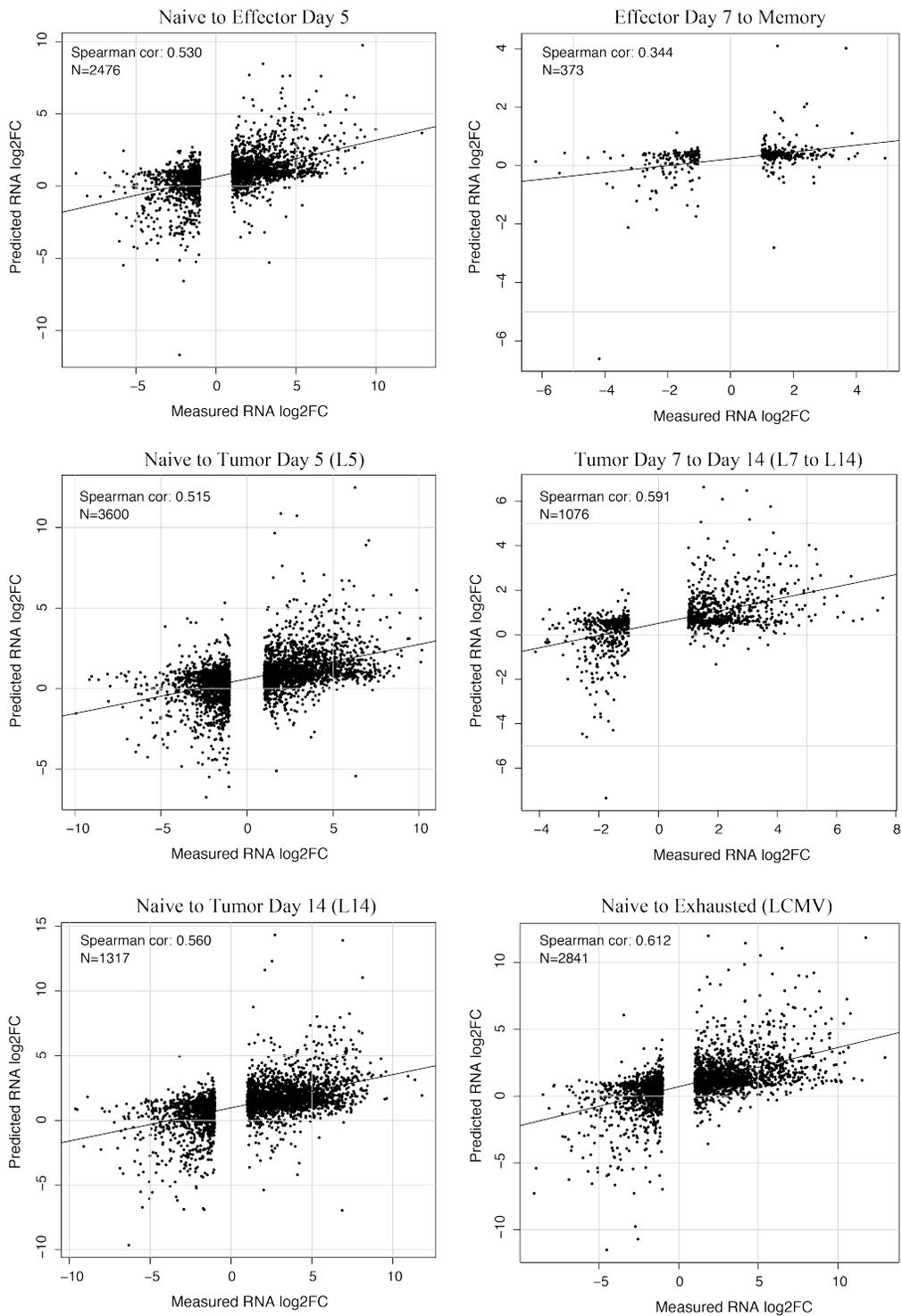


Figure 4.2: Model performance scatterplots. Scatterplots depicting measured RNA log₂ fold change vs predicted RNA log₂ fold change for various cell type transitions

Table 4.2: Model performance across six cell type transitions

Spearman (rank) correlations of measured vs predicted RNA fold change in six cell type transitions. First column: Correlation calculated using all differentially expressed genes. Second column: Correlation calculated using only high-complexity differentially expressed genes (genes with 7 or more assigned peaks).

Cell type transition	Spearman corr. (all genes)	Spearman corr. (high-complexity genes)
N to E5	0.534	0.610
E7 to M	0.319	0.430
N to L5	0.511	0.569
N to L14	0.630	0.630
L7 to L14	0.592	0.639
N to Exhausted (LCMV)	0.615	0.647

Next, we looked at the learned weights from the model to determine the relative importance of each transcription factor in the prediction of gene expression changes. By multiplying the learned TF weight vector (W) by the binding scores derived from cisBP and the ATAC-seq data, we are able to determine which TFs are best able to explain changes in gene expression for each cell type transition. As a positive control, we were able to recover several TFs known to be critical in T cell activation from naïve to effector states: Bach1, Eomes, Nfat family transcription factors, and Tbx21 (Tbet). We also found that closing of Lef1/Tcf family binding sites explained the lower expression of Naïve-specific genes in effector cells (**Figure 4.3 (left), Table 4.3**).

In contrast to normal T cell activation, Klf family proteins seemed to play a stronger role in the transitions from naïve to tumor-specific dysfunction and to LCMV-induced exhaustion. Both of these transitions showed many members of the Klf family explaining gene expression changes in both the positive and negative directions. Additionally, neither of the dysfunctional transitions showed Tbx21 (Tbet) to be a

major influencer. The activation of Nfat family members also appeared to be muted in the transition to LCMV-induced exhaustion, in contrast to the hyper-activation of Nfat in the tumor-specific dysfunction transition. More analysis and experimental follow-up is needed to confirm these findings.

Based on this analysis, it appears that although the chromatin landscapes of tumor-specific dysfunctional T cells and LCMV-induced exhausted T cells appear similar (**Figure 4.1**), there are differences in the underlying regulatory mechanisms controlling the cell state changes (**Figure 4.3**).

Table 4.3: Influential transcription factors identified by model

Summary of Figure 4.3 depicting learned weights for each transcription factor weighted by the binding score derived from cisBP. Here are shown transcription factors that explain up- or down- regulation of genes in each of the three cell type transitions analyzed.

	Up-regulation	Down-regulation
Naïve to E5	Bach1, Batf, Eomes, Fos/Jun, Klf, Nfat, Nfkb2, Pbx2, Runx1/3, Smarcc1, Tbx21, Trp53	Klf, Lef1/Tcf, Pbx2, Trp53
Naïve to L5	Bach1, Eomes, Fosb, Klf, Nfatc1, Nfkb1, Klf, Pbx2	Bach2, Klf6, Nfkb1/2, Pbx2, Tcf, Trp53
Naïve to Exhausted	Bach1, Eomes, Fos, Fox, Klf, Lef1, Nfatc2, Prdm1	Klf, Lef1, Sox4, Tcf, Trp53

Figure 4.3: Heatmaps of learned transcription factor weights scaled by cisBP-derived binding terms.

Each heatmap depicts the relative importance of each transcription factor in the up- or down-regulation of genes during cell type transitions. (left) Naïve to effector day 5 (E5) (center) Naïve to tumor-associated dysfunction day 5 (L5) (right) Naïve to LCMV-induced exhaustion. Columns represent up-regulated (orange) or down-regulated (blue) genes in each cell type transition. Rows represent transcription factor binding sites that open (orange) or close (blue) in each transition. Rows are additionally annotated as either promoter sites (pink) or enhancer sites (teal). Red indicates that the transcription factor contributed positively to gene expression (up-regulation) and blue indicates the opposite (down-regulation). Selected transcription factors are highlighted to the right of each heatmap.

CHAPTER 5

CONCLUSIONS AND FUTURE DIRECTIONS

5.1 Discussion

The question of gene regulation is a complex and varied one. In this work I describe two studies that investigate different angles of gene regulation. In chapter 2 I describe SplashRNA, a cascaded SVM classifier trained to predict potent shRNA reagents. In chapter 3 I present an in-depth analysis of chromatin accessibility in CD8⁺ T cells and how this accessibility changes and potentially regulates cell state changes. An extension of this analysis was presented in chapter 4, in which the relationship between accessible chromatin, predicted transcription factor binding sites, and the resulting changes in gene expression were modeled using a generalized linear model.

5.2 shRNA potency prediction

SplashRNA was developed to more accurately predict potent shRNA molecules using only sequence information. It is comprised of a linear combination of two SVM classifiers, one trained on extensive miR-30 shRNA backbone data, and the second trained on a smaller amount of the newer miR-E backbone data. This combination allowed the classifier to learn basic rules of shRNA potency from the larger but slightly out-of-date miR-30 data and then alter those rules using the sparse but newer miR-E data. This combination led to unprecedented prediction performance in novel prediction tasks relative to existing methods (shRNA with a SplashRNA score >1 have an 80% probability of strong knockdown and over 80% of human and mouse genes have at least 5 predicted shRNAs with scores of 1 or greater). The SplashRNA project also included

an extensive curation of the transcribed sequences in human and mouse genomes in order to target the relevant transcript space in both species. Only sequences transcribed in all RNA isoforms of a gene were used as input and poly-adenylation libraries were used to determine the most proximal 3' polyadenylation site for each gene in order to prevent targeting a gene after this cleavage event. These features were incorporated with the SplashRNA algorithm into an open-source online tool that can be found at splashrna.mskcc.org. A docker instance of the algorithm is also available for download (<https://hub.docker.com/r/lfairchild/splashrna-docker/>).

Though this approach of cascading classifiers was developed for face detection by sequentially detecting potential face components and exiting calculation early if no components are found, we applied the concept here for shRNA prediction. This method of cascading SVM classifiers could easily be extended to newer biological developments such as CRISPR/cas9 reagent prediction. Although there is currently not a large pool of CRISPR training data and the specific targeting rules to vary from those used in shRNA/siRNA, the overall problem is similar. More broadly, any situation with a shift in the target prediction (e.g. miR-30 to miR-E prediction) may benefit from this type of approach, as the cascade allows for the classification rules to be refined at each step.

Additionally, although a strength of this classifier lies in its ability to predict shRNA potency using only sequence features, in some circumstances it may be advantageous to add other non-sequence features to the model, potentially increasing performance. Potential non-sequence features would include RNA structure prediction to locate regions of the target mRNA that could potentially be bound, RNA modifications including methylation which could conceivably affect shRNA/RISC interaction with the target mRNA, and thermodynamic stability of the shRNA/mRNA complex.

Innovations in shRNA technology continue to push the boundary of what is possible with shRNA. Constructs are in development that can deliver several shRNA at once, allowing for the knock down of multiple genes simultaneously. Since multiple shRNA molecules are being delivered, a heavier burden is being placed on the cell's machinery. Early experiments suggest that SplashRNA predictions continue to perform well in this setting, but some optimization may improve performance.

5.3 T cell epigenetic modeling

Although the immune system and CD8+ T cells have been studied extensively, it is still not clear what causes them to become inappropriately nonfunctional, for example in the presence of tumor cells. In chapter 3 we addressed this question by molecularly profiling CD8+ T cells in acute infection and malignant tumor environments using both ATAC-seq and RNA-seq. This in-depth sequence information allowed us to characterize the T cells to an extent not previously possible and to determine chromatin states corresponding to normal and dysfunctional T cell states. Using RNA-seq data, we were able to identify cell-surface markers that could reliably differentiate between TSTs extracted from the premalignant environment that could be functionally rescued from those that were in a fixed dysfunctional state. This type of testing could potentially be adapted to patients to determine whether they are candidates for checkpoint blockade therapy; meaning that their T cells could potentially be reinvigorated by an anti-PD1 or anti-CTLA4 antibody.

In addition to determining cell-surface markers that differentiate TSTs, we analyzed the accessibility of predicted transcription factor binding sites to determine which transcription factors were most likely involved in the transitions from naïve to early dysfunctional cells and from early (plastic) dysfunction to late (fixed) dysfunction.

This analysis revealed that NFAT family member binding sites became significantly more accessible in early dysfunctional cells relative to their naïve predecessors, and that binding sites for TCF family members (including LEF1) closed during the transition from early to late dysfunction, correlating with the decline in TCF1 protein levels. Furthermore, we were able to validate the roles of these two predicted transcription factor families in the progression to T cell dysfunction using an *in vivo* pharmacologic strategy. Pharmacologically down-regulating NFAT activity and activating the WNT/TCF1 pathway led to significantly more cells remaining reprogrammable relative to controls after being exposed to the dysfunction-inducing environment, indicating that these transcription factor families play a role in regulating this transition to dysfunction.

Although we did identify transcription factors that play a role in the progression to dysfunction and we were able to slow the progression to dysfunction by targeting these TF families, we do not know if the result we observed was only at the level of phenotype (restoration of effector cytokines) or if the treatment also slowed the chromatin remodeling to the dysfunctional signature. This would give an indication as to whether the drug treatment was able to fully protect some cells from progressing to dysfunction, or if it was only delaying this progression.

Another question that remains unanswered is whether the transcription factors are acting primarily through a relatively small set of “key” genes, or if their effect is truly global. In the case where transcription factors are modulating a small number of genes, the binding sites for these transcription factors could be individually modulated using CRISPR, as was done for a potential PD1-regulating site in Sen & Kaminski et al. (2016)⁷³, in order to determine the precise effect of each transcription factor. More likely however, is that the effects of these transcription factors form a complicated genome-wide network of signals that would be impossible to elucidate experimentally.

In chapter 4, I attempted to take this type of transcription factor binding analysis one step farther by directly modeling the impact of each expressed transcription factor on gene expression during a cell type transition. By doing this, I attempted to determine the relative importance of each transcription factor in orchestrating that cell type transition. This type of analysis more directly addresses the molecular underpinnings of these cell state changes and may point toward potential treatment paths that have not been explored. Although a model for each cell state change is technically true, having one model that could explain all the cell type changes, for example by learning in a multitask setting, would be a cleaner solution. This strategy could learn a universal transcription factor weighting and then a task (or cell type transition) specific weighting that would modify the transcription factor importance weights for each individual cell type transition. Additionally, allowing the model to learn from all cell type transitions would likely increase performance as the training data would be increased.

The method used to match transcription factors to target genes, using predicted binding scores derived from cisBP, can likely be improved upon to increase overall model performance. It is not feasible to collect *in vivo* binding data for all transcription factors in the cell type of interest, but perhaps creating an atlas of binding sites for transcription factors profiled in the ENCODE project (encyclopedia of DNA elements)¹⁰⁷ in any cell type would provide a more realistic view of the landscape of binding events for some transcription factors.

An additional challenge is mapping regions of accessible chromatin genes they are regulating. Currently, we map each chromatin element to its nearest gene but this assumes that regulatory elements are always closely positioned to their target genes. Having DNA looping information such as Hi-C or Hi-ChIP data would be valuable in determining topologically associating domains (TADs), or regions that are likely to be co-regulated. However, since this data would be difficult to generate in many rare cell

populations, it may be more feasible to reassign peaks associated with high-error genes, as proposed in Gonzalez & Setty et al (2015)¹⁰⁸.

5.4 Conclusion

The two studies included in this document describe forays into different realms of biology, but at their hearts they are both questions of gene regulation that are being answered predominantly in a computational manner. In chapter 2 I described SplashRNA, a machine learning method to predict potent shRNAs from sequence alone. This method allows for higher efficiency in the targeted knockdown of RNA transcripts in the lab – a critical tool in the study of gene regulation. In chapter 3 I described a computational analysis of chromatin accessibility and gene expression data in tumor-specific T cells. This study seeks to uncover the principles guiding gene regulation in this setting and would have been impossible without computational investigation. Together, these two studies show the breadth and importance of computational analysis and innovation in the biological sciences.

BIBLIOGRAPHY

1. Harrow, J. *et al.* GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Research* **22**, 1760–1774 (2012).
2. Mudge, J. M. & Harrow, J. Creating reference gene annotation for the mouse C57BL6/J genome assembly. *Mammalian Genome* **26**, 366–378 (2015).
3. Hillier, L. W. *et al.* Genomics in *C. elegans*: so many genes, such a little worm. *Genome Research* **15**, 1651–1660 (2005).
4. Engel, S. R. *et al.* The reference genome sequence of *Saccharomyces cerevisiae*: then and now. *G3 (Bethesda)* **4**, 389–398 (2014).
5. Levine, M. & Tjian, R. Transcription regulation and animal diversity. *Nature* **424**, 147–151 (2003).
6. Fire, A. *et al.* Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* **391**, 806–811 (1998).
7. Fellmann, C. & Lowe, S. W. Stable RNA interference rules for silencing. *Nature Cell Biology* **16**, 10–18 (2014).
8. Kamola, P. J., Nakano, Y., Takahashi, T., Wilson, P. A. & Ui-Tei, K. The siRNA Non-seed Region and Its Target Sequences Are Auxiliary Determinants of Off-Target Effects. *PLoS Comput. Biol.* **11**, e1004656 (2015).
9. Pelosof, R. *et al.* Prediction of potent shRNAs with a sequential classification algorithm. *Nat Biotech* **35**, 350–353 (2017).
10. Fellmann, C., Zuber, J., McJunkin, K. & Chang, K. Functional identification of optimized RNAi triggers using a massively parallel sensor assay. *Mol. Cell* (2011).
11. Yuan, T. L. *et al.* Development of siRNA Payloads to Target KRAS-Mutant Cancer. *Cancer Discovery* **4**, 1182–1197 (2014).
12. Knott, S. R. V. *et al.* A computational algorithm to predict shRNA potency. *Mol. Cell* **56**, 796–807 (2014).
13. Fellmann, C., Hoffmann, T., Sridhar, V. & Hopfgartner, B. An optimized microRNA backbone for effective single-copy RNAi. *Cell Reports* (2013).
14. Auyeung, V. C., Ulitsky, I., McGeary, S. E. & Bartel, D. P. Beyond secondary structure: primary-sequence determinants license pri-miRNA hairpins for processing. *Cell* **152**, 844–858 (2013).
15. Khvorova, A., Reynolds, A. & Jayasena, S. D. Functional siRNAs and miRNAs Exhibit Strand Bias - ScienceDirect. *Cell* **115**, 209–216 (2003).
16. Reynolds, A. *et al.* Rational siRNA design for RNA interference. *Nat. Biotechnol.* **22**, 326–330 (2004).
17. Schwarz, D. S. *et al.* Asymmetry in the assembly of the RNAi enzyme complex. *Cell* **115**, 199–208 (2003).
18. Huesken, D. *et al.* Design of a genome-wide siRNA library using an artificial neural network. *Nat. Biotechnol.* **23**, 995–1001 (2005).
19. Vert, J.-P., Foveau, N., Lajaunie, C. & Vandenbrouck, Y. An accurate and interpretable model for siRNA efficacy prediction. *BMC Bioinformatics* **7**, 520

- (2006).
20. Saetrom, P. & Snøve, O. A comparison of siRNA efficacy predictors. *Biochem. Biophys. Res. Commun.* **321**, 247–253 (2004).
 21. Filhol, O. *et al.* DSIR: Assessing the Design of Highly Potent siRNA by Testing a Set of Cancer-Relevant Target Genes. *PLoS ONE* **7**, e48057 (2012).
 22. Taxman, D. J. *et al.* Criteria for effective design, construction, and gene knockdown by shRNA vectors. *BMC Biotechnol.* **6**, 7 (2006).
 23. Matveeva, O. V., Nazipova, N. V., Ogurtsov, A. Y. & Shabalina, S. A. Optimized models for design of efficient miR30-based shRNAs. *Front. Genet.* **3**, (2012).
 24. Smith-Garvin, J. E., Koretzky, G. A. & Jordan, M. S. T Cell Activation. *Annu. Rev. Immunol.* **27**, 591–619 (2009).
 25. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009).
 26. Fritsch, E. F., Hacoheh, N. & Wu, C. J. Personal neoantigen cancer vaccines. *OncoImmunology* **3**, e29311 (2014).
 27. Vinay, D. S. *et al.* Immune evasion in cancer: Mechanistic basis and therapeutic strategies. *Seminars in Cancer Biology* **35**, S185–S198 (2015).
 28. Füreder, T. *et al.* Review of cancer treatment with immune checkpoint inhibitors. *Wiener klinische Wochenschrift* 1–7 (2018). doi:10.1007/s00508-017-1285-9
 29. Wherry, E. J. T cell exhaustion. *Nat. Immunol.* **12**, 492–499 (2011).
 30. Schietinger, A. *et al.* Tumor-Specific T Cell Dysfunction Is a Dynamic Antigen-Driven Differentiation Program Initiated Early during Tumorigenesis. *Immunity* **45**, 389–401 (2016).
 31. Kornberg, R. D. & Lorch, Y. Chromatin Structure and Transcription. *Annual Review of Cell Biology* **8**, 563–587 (1992).
 32. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. 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).
 33. Richmond, T. J. & Davey, C. A. The structure of DNA in the nucleosome core. *Nature* **423**, 145–150 (2003).
 34. Guda, S. *et al.* miRNA-embedded shRNAs for Lineage-specific BCL11A Knockdown and Hemoglobin F Induction. *Mol. Ther.* **23**, 1465–1474 (2015).
 35. Grimm, D. *et al.* Fatality in mice due to oversaturation of cellular microRNA/short hairpin RNA pathways. *Nature* **441**, 537–541 (2006).
 36. McBride, J. L., Boudreau, R. L. & Harper, S. Q. Artificial miRNAs mitigate shRNA-mediated toxicity in the brain: implications for the therapeutic development of RNAi. in (2008).
 37. Baek, S. T. *et al.* Off-target effect of doublecortin family shRNA on neuronal migration associated with endogenous microRNA dysregulation. *Neuron* **82**, 1255–1262 (2014).
 38. Zuber, J. *et al.* Toolkit for evaluating genes required for proliferation and survival using tetracycline-regulated RNAi. *Nat Biotech* **29**, 79–83 (2011).

39. Gu, S. *et al.* The loop position of shRNAs and pre-miRNAs is critical for the accuracy of dicer processing in vivo. *Cell* **151**, 900–911 (2012).
40. Watanabe, C., Cuellar, T. L. & Haley, B. Quantitative evaluation of first, second, and third generation hairpin systems reveals the limit of mammalian vector-based RNAi. *RNA Biol*
41. Viola, P. & Jones, M. Rapid object detection using a boosted cascade of simple features. ... *Vision and Pattern Recognition* (2001).
42. Pelossof, R. A. Rapid Learning with Stochastic Focus of Attention Raphael A. Pelossof. *COLUMBIA UNIVERSITY* 1–98 (2011).
43. Leslie, C. S., Eskin, E. & Noble, W. S. The spectrum kernel: A string kernel for SVM protein classification. *Pacific symposium on ...* (2002).
44. Sonnenburg, S., bf, G. R. & Rieck, K. Large Scale Learning with String Kernels. *researchgate.net* 73–104
45. Kampmann, M. *et al.* Next-generation libraries for robust RNA interference-based genome-wide screens. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E3384–91 (2015).
46. Morgens, D. W., Deans, R. M., Li, A. & Bassik, M. C. Systematic comparison of CRISPR/Cas9 and RNAi screens for essential genes. *Nat Biotech* **34**, 634–636 (2016).
47. Kampmann, M., Bassik, M. C. & Weissman, J. S. Integrated platform for genome-wide screening and construction of high-density genetic interaction maps in mammalian cells. *Proc. Natl. Acad. Sci. U.S.A.* **110**, E2317–26 (2013).
48. Hart, T., Brown, K. R., Sircoulomb, F., Rottapel, R. & Moffat, J. Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. *Mol. Syst. Biol.* **10**, 733–733 (2014).
49. Spies, N., Burge, C. B. & Bartel, D. P. 3' UTR-isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts. *Genome Research* **23**, 2078–2090 (2013).
50. Derti, A. *et al.* A quantitative atlas of polyadenylation in five mammals. *Genome Research* **22**, 1173–1183 (2012).
51. Lianoglou, S., Garg, V., Yang, J. L., Leslie, C. S. & Mayr, C. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes & Development* **27**, 2380–2396 (2013).
52. Yi, R., Doehle, B. P., Qin, Y., Macara, I. G. & Cullen, B. R. Overexpression of exportin 5 enhances RNA interference mediated by short hairpin RNAs and microRNAs. *RNA* **11**, 220–226 (2005).
53. Boudreau, R. L., Martins, I. & Davidson, B. L. Artificial microRNAs as siRNA shuttles: improved safety as compared to shRNAs in vitro and in vivo. *Mol. Ther.* **17**, 169–175 (2009).
54. Sigoillot, F. D. *et al.* A bioinformatics method identifies prominent off-targeted transcripts in RNAi screens. *Nat. Methods* **9**, 363–366 (2012).
55. Philip, M. *et al.* Chromatin states define tumour-specific T cell dysfunction and reprogramming. *Nature* **545**, 452–456 (2017).
56. Hellström, I., Hellström, K. E., Pierce, G. E. & Yang, J. P. S. Cellular and Humoral immunity to Different Types of Human Neoplasms. *Nature* **220**,

- 1352–1354 (1968).
57. Khalil, D. N., Smith, E. L., Brentjens, R. J. & Wolchok, J. D. The future of cancer treatment: immunomodulation, CARs and combination immunotherapy. *Nature Reviews Clinical Oncology* **13**, 273–290 (2016).
 58. Snyder, A. *et al.* Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma. *N Engl J Med* **371**, 2189–2199 (2014).
 59. Kelderman, S., Schumacher, T. N. M. & Haanen, J. B. A. G. Acquired and intrinsic resistance in cancer immunotherapy. *Molecular Oncology* **8**, 1132–1139 (2014).
 60. Rizvi, N. A. *et al.* Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* **348**, 124–128 (2015).
 61. Zhang, J. A., Mortazavi, A., Williams, B. A., Wold, B. J. & Rothenberg, E. V. Dynamic transformations of genome-wide epigenetic marking and transcriptional control establish T cell identity. *Cell* **149**, 467–482 (2012).
 62. Scharer, C. D., Barwick, B. G., Youngblood, B. A., Ahmed, R. & Boss, J. M. Global DNA Methylation Remodeling Accompanies CD8 T Cell Effector Function. *The Journal of Immunology* **191**, 3419–3429 (2013).
 63. Russ, B. E., Olshanksy, M., Smallwood, H. S., Li, J. & Denton, A. E. Distinct epigenetic signatures delineate transcriptional programs during virus-specific CD8+ T cell differentiation. *Immunity* (2014).
 64. Shih, H.-Y. *et al.* Developmental Acquisition of Regulomes Underlies Innate Lymphoid Cell Functionality. *Cell* **165**, 1120–1133 (2016).
 65. Staveley-O'Carroll, K. *et al.* In Vivo Ligation of CD40 Enhances Priming Against the Endogenous Tumor Antigen and Promotes CD8+ T Cell Effector Function in SV40 T Antigen Transgenic Mice. *The Journal of Immunology* **171**, 697–707 (2003).
 66. Brockstedt, D. G. *et al.* Listeria-based cancer vaccines that segregate immunogenicity from toxicity. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 13832–13837 (2004).
 67. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* **40**, 4288–4297 (2012).
 68. Araki, Y., Fann, M., Wersto, R. & Weng, N.-P. Histone Acetylation Facilitates Rapid and Robust Memory CD8 T Cell Response through Differential Expression of Effector Molecules (Eomesodermin and Its Targets: Perforin and Granzyme B). *Journal of immunology (Baltimore, Md. : 1950)* **180**, 8102 (2008).
 69. Denton, A. E., Russ, B. E., Doherty, P. C., Rao, S. & Turner, S. J. Differentiation-dependent functional and epigenetic landscapes for cytokine genes in virus-specific CD8+ T cells. *pnas.org*
 70. Best, J. A. *et al.* Transcriptional insights into the CD8+ T cell response to infection and memory T cell formation. *Nat. Immunol.* **14**, 404–412 (2013).
 71. Cuylen, S. *et al.* Ki-67 acts as a biological surfactant to disperse mitotic chromosomes. *Nature* **535**, 308–312 (2016).
 72. Kaech, S. M. & Cui, W. Transcriptional control of effector and memory CD8+

- T cell differentiation. *Nat Rev Immunol* **12**, 749–761 (2012).
73. Sen, D. R. *et al.* The epigenetic landscape of T cell exhaustion. *Science* **354**, 1165–1169 (2016).
 74. Pauken, K. E. *et al.* Epigenetic stability of exhausted T cells limits durability of reinvigoration by PD-1 blockade. *Science* **354**, 1160–1165 (2016).
 75. Scott-Browne, J. P. *et al.* Dynamic Changes in Chromatin Accessibility Occur in CD8+ T Cells Responding to Viral Infection. *Immunity* **45**, 1327–1340 (2016).
 76. Macian, F. NFAT proteins: key regulators of T-cell development and function. *Nat Rev Immunol* **5**, 472–484 (2005).
 77. Martinez, G. J. *et al.* The transcription factor NFAT promotes exhaustion of activated CD8+ T cells. *Immunity* **42**, 265–278 (2015).
 78. Teague, R. M. *et al.* Interleukin-15 rescues tolerant CD8+ T cells for use in adoptive immunotherapy of established tumors. *Nat Med* **12**, 335–341 (2006).
 79. Li, Y. *et al.* MART-1–Specific Melanoma Tumor-Infiltrating Lymphocytes Maintaining CD28 Expression Have Improved Survival and Expansion Capability Following Antigenic Restimulation In Vitro. *The Journal of Immunology* **184**, 452–465 (2009).
 80. Flanagan, W. M., Corthésy, B., Bram, R. J. & Crabtree, G. R. Nuclear association of a T-cell transcription factor blocked by FK-506 and cyclosporin A. *Nature* **352**, 803–807 (1991).
 81. Jain, J. *et al.* The T-cell transcription factor NFATp is a substrate for calcineurin and interacts with Fos and Jun. *Nature* **365**, 352–355 (1993).
 82. Gattinoni, L. *et al.* Wnt signaling arrests effector T cell differentiation and generates CD8+ memory stem cells. *Nat Med* **15**, 808–813 (2009).
 83. Schietinger, A., Delrow, J. J., Basom, R. S., Blattman, J. N. & Greenberg, P. D. Rescued Tolerant CD8 T Cells Are Preprogrammed to Reestablish the Tolerant State. *Science* **335**, 723–727 (2012).
 84. Waugh, K. A. *et al.* Molecular Profile of Tumor-Specific CD8 +T Cell Hypofunction in a Transplantable Murine Cancer Model. *The Journal of Immunology* **197**, 1477–1488 (2016).
 85. Krueger, F. Trim Galore! (2017).
 86. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet j.* **17**, 10 (2011).
 87. Andrews, S. FastQC. (2015).
 88. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
 89. Adey, A. *et al.* Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol* **11**, R119 (2010).
 90. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).
 91. Corwin M Zigler, T. R. B. THE POTENTIAL FOR BIAS IN PRINCIPAL CAUSAL EFFECT ESTIMATION WHEN TREATMENT RECEIVED DEPENDS ON A KEY COVARIATE. *The annals of applied statistics* **5**, 1876–1892 (2011).

92. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol* **11**, R106 (2010).
93. Lawrence, M. *et al.* Software for Computing and Annotating Genomic Ranges. *PLoS Comput. Biol.* **9**, e1003118 (2013).
94. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
95. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat Biotech* **29**, 24–26 (2011).
96. Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202–W208 (2009).
97. Weirauch, M. T. *et al.* Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. *Cell* **158**, 1431–1443 (2014).
98. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
99. Hinrichs, A. S. *et al.* The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* **34**, D590–D598 (2006).
100. Denas, O. *et al.* Genome-wide comparative analysis reveals human-mouse regulatory landscape and evolution. *BMC Genomics* **16**, 87 (2015).
101. Chen, A.-J. *et al.* STAR RNA-binding protein Quaking suppresses cancer via stabilization of specific miRNA. *Genes & Development* **26**, 1459–1472 (2012).
102. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotech* **28**, 495–501 (2010).
103. Consortium, T. G. O. Gene Ontology Consortium: going forward. *Nucleic Acids Res.* **43**, D1049–D1056
104. Xing, Y. & Hogquist, K. A. T-Cell Tolerance: Central and Peripheral. *Cold Spring Harbor Perspectives in Biology* **4**, a006957–a006957 (2012).
105. Lara-Astiaso, D. *et al.* Immunogenetics. Chromatin state dynamics during blood formation. *Science* **345**, 943–949 (2014).
106. Pauken, K. E. *et al.* Epigenetic stability of exhausted T cells limits durability of reinvigoration by PD-1 blockade. *Science* **354**, 1160–1165 (2016).
107. ENCODE Project Consortium, T. E. P. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **488**, 57–74 (2012).
108. Gonzalez, A. J., Setty, M. & Leslie, C. S. Early enhancer establishment and regulatory locus complexity shape transcriptional programs in hematopoietic differentiation. *Nat. Genet.* 1–14 (2015). doi:10.1038/ng.3402