Retroviral Dysregulation of an Immune Chromatin Regulator

by

Adam Kenneth Dziulko

B.S., Iowa State University, 2015

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado at Boulder in partial fulfillment

of the requirement for the degree of

Doctor of Philosophy

Molecular, Cellular, and Developmental Biology

2024

Committee Members:

Dr. Edward Chuong

Dr. Justin Brumbaugh

Dr. Robin Dowell

Dr. Corrella Detweiler

Dr. Dylan Taatjes

Abstract

Dziulko, Adam Kenneth (Ph.D., Molecular, Cellular, and Developmental Biology)

Retroviral Dysregulation of an Immune Chromatin Regulator

Thesis directed by Assistant Professor Edward Chuong

Transposable elements (TEs) are an abundant source of gene regulatory sequence in the genome. Some TEs have been co-opted by the host genome to serve beneficial regulatory functions, however some TEs have been found to activate disease. Senescence, a pro-inflammatory cell state, is an attractive disease model given its implication in disease. Using an oncogene-induced senescence (OIS) model, I asked whether TEs have any gene regulatory effects that contribute to senescence pro-inflammatory milieu. There were many gene candidates that could potentially be regulated by a TE in the OIS model, however, I settled on Speckled Protein 140 (SP140) from observing a potential novel transcript being produced out of its normal immune modulatory context.

SP140 is a chromatin reader with critical roles regulating immune cell transcriptional programs, and SP140 splice variants are associated with immune diseases including Crohn's disease, multiple sclerosis, and chronic lymphocytic leukemia. SP140 expression is currently thought to be restricted to immune cells. However, by analyzing human transcriptomic datasets from a wide range of normal and cancer cell types, we found recurrent cancer-specific expression of SP140, driven by an alternative promoter derived from an intronic endogenous retrovirus (ERV). The ERV belongs to the primate-specific LTR8B family and is regulated by oncogenic mitogen-activated protein kinase (MAPK) signaling. The ERV drives expression of multiple cancer-specific isoforms, including a nearly full-length isoform that retains all the

functional domains of the full-length canonical isoform and is also localized within the nucleus, consistent with a role in chromatin regulation. In a fibrosarcoma cell line, silencing the cancer-specific ERV promoter of SP140 resulted in phenotypic changes less suited for cancer cells - increased sensitivity to interferon-mediated cytotoxicity and increased regulation of a tumorigenic transcription factor, nuclear factor kappa-B (NF-kB). Our findings implicate aberrant ERV-mediated SP140 expression as a novel mechanism contributing to immune gene dysregulation in a wide range of cancer cells.

ACKNOWLEDGMENTS

I thank Dr. Chuong for his guidance and mentorship. I thank the current and past members of the Chuong Lab for feedback and input, as well as their support. I thank the Interdisciplinary quantitative biology program for their support. I thank my collaborators Dr. Kristen Witt and Dr. Russel Vance. I thank Dr. Holly Allen for help with experiments and editing drafts. I thank the University of Colorado Anschutz Cancer Center Genomics Core (RRID: SCR_021984; P30CA046934) where sequencing was conducted. I thank the University of Colorado Genomics Shared Resource and BioFrontiers Computing core for technical support during this study. E.B.C. was supported by the National Institutes of Health (1R35GM128822), the Alfred P. Sloan Foundation, the David and Lucile Packard Foundation, the Boettcher foundation, and a Discovery Boost Grant, DBG-23-1155983-01-DMC, from the American Cancer Society (Grant DOI #: https://doi.org/10.53354/ACS.DBG-23-1155983-01-DMC.pc.gr.175433).

Contents

Figures

Chapter I: Introduction

TEs in the human genome

The 32-year path to sequencing and mapping the human genome started in 1990, where it took 11 years to initially sequence and analyze ~2.4 billion nucleotides (~79% of total human genome) with 1 billion nucleotides (~33% of total human genome) being mapped [1]. Three years later in 2004, 2.85 billion nucleotides (~93% of total human genome) was sequenced and mapped [2], then 18 years later in 2022 the full 3.055 billion nucleotides of the human genome was sequenced and mapped [3]. From these efforts, the genome is broken down into groups with approximately: 2% protein-coding genes, 26% introns, 3% simple sequence repeats, 5% segmental duplications, 6-8% miscellaneous heterochromatin, 10-12% miscellaneous unique sequences, and 46% transposable elements (TEs) [1,3,4]. TEs take up nearly half of the genome and possibly more (up to 66%), with degraded/remnant TEs currently classified as repeats [5]. TEs account for ~5 million loci mostly at 100 to 10,000 base pairs in length in the human genome. Here, I will cover these abundant sequences and their implications in the genome.

TEs are selfish DNA sequences that can move, or "transpose", within the genome. They are sometimes referred to as "jumping genes" because of their ability to change their position. In humans and other organisms, transposable elements have been present for a very long time in evolutionary terms with the vast majority of insertions estimated to be between 5 and 60 million years old with humans having ~25-30% TEs unique to them. The presence of TEs specific to humans is a result of ancient viral infections, horizontal gene transfer, differential expansion, genomic isolation, and positive selection [6]. These elements have played significant roles in

shaping the human genome, contributing to both genetic diversity and regulatory innovation. The fraction of the genome occupied by TEs does not correlate with organismal complexity with some fungal species being as low as <1% of TEs in their genome [7], the western honey bee containing <5% TEs [8], mouse at ~38% TEs [9], the Gray short-tailed opossum at ~52% TEs [10], and corn being as high as >85% TEs [11]. TE abundance, however, correlates more strongly with genome size with genome sizes at ~36.5 million base pairs, ~236 million base pairs, ~2.5 billion base pairs, ~3.6 billion base pairs, and ~2.4 billion base pairs, respectively to the previous sentence.

Basic mechanisms of TE integration and replication

TEs can integrate somewhat randomly within the genome, however, some factors influence their integration patterns. These influences are often guided by a balancing act of facilitating future propagation while mitigating deleterious effects on host genome structure and function. Many TEs have evolved mechanisms to target specific loci where their insertions are less detrimental to the host and favorable for their propagation [12]. Further, Natural selection and genetic drift shape the distribution of TEs [13].

TEs come in a variety of forms based on their structure, mechanism of transposition, and the presence or absence of certain enzymatic activities. TEs integrate and replicate through distinct mechanisms depending on their class. DNA transposons use a cut-and-paste mechanism involving a transposase enzyme, while retrotransposons utilize a copy-and-paste mechanism involving reverse transcription. Long Terminal Repeat (LTR) retrotransposons produce cDNA in cytoplasmic particles before integration, whereas non-LTR retrotransposons use target-primed

reverse transcription to integrate directly at the target site. These mechanisms allow TEs to proliferate within genomes, contributing to genetic diversity and evolution [14–16].

Humans have a total of ~1,103 TE families and subfamilies which are classified as DNA sequences longer than 80 base pairs that share at least 80% sequence identity over 80% of their length [17] traced as descendants of a single ancestral unit. It is generally accepted that the majority of major TE families stem from a limited set of actively expressed copies referred to as "source" or "founder" genes, generating distinct families of "repetitive elements" [18]. Subfamilies are subsets within these families that exhibit additional variations or evolutionary divergence.

These (sub)families can be divided into two major classes based on their mechanism of transposition, and each class can be subdivided into subclasses based on the mechanism of chromosomal integration. Class I TEs (retrotransposons) transpose via an RNA intermediate using a "copy and paste" mechanism. Class I TEs include LTRs, Long INterspersed Elements (LINEs), and Short INterspersed Elements (SINEs). LTRs take up 8% of the human genome and ~150 families have direct repeats at their ends and encode proteins necessary for their transposition. LINEs take up 20% of the genome and encode a reverse transcriptase and an endonuclease that can transpose autonomously. SINEs take up 13% of the genome and are non-autonomous elements that rely on LINEs for their transposition. Class II Transposable Elements (DNA transposons; 3% of the genome) transpose directly as DNA either through a cut-and-paste or replicative method.

Impact of TEs on human genome function

TEs can play a crucial role in genome evolution and function, affecting human genome regulation in several ways: inserting themselves into or near genes which can disrupt gene function directly by breaking the gene sequence or altering its regulatory regions; inserting themselves within or near genes which can lead to the creation of new exon-intron structures resulting in alternative splicing events; causing chromosomal rearrangements such as deletions, duplications, and inversions; affecting the epigenetic state of their surrounding genomic regions; and carrying their own regulatory sequences, such as promoters or enhancers, that can influence the expression of nearby genes [14,15,19,20].

Though TEs are named and classified after their 'jumping' characteristics, only ~35-40 (sub)families (<.05% of the genome) remain actively mobile in the human genome [21]. Most human TEs are 'fixed' meaning they have lost their ability to move around in the genome due to consequences of evolutionary processes such as the lack of selective pressure to maintain active transposition, mutation accumulation, and host defense mechanisms.

To fight back against TEs negatively affecting cell fitness, host organisms have evolved intricate defense mechanisms to regulate TE activity. One such defense system involves epigenetic modifications, where TEs are marked with repressive marks such as DNA methylation and histone modifications [22,23]. These marks can lead to transcriptional silencing of TEs, preventing their expression and transposition. Additionally, host cells employ RNA interference pathways including RNA-induced Silencing Complex and Piwi-interacting RNA pathway proteins to target TE transcripts and inhibit their mobility [24–27]. Another defense mechanism involves Kruppel-associated box zinc finger proteins (KRAB-ZFPs). KRAB-ZFPs evolve alongside with TEs in an 'evolutionary arms race' to recognize and bind TEs that are

evolving [28–30]. Through these defense strategies, host organisms can mitigate the harmful effects of TEs, maintaining genomic integrity and stability over evolutionary time.

TE exaptation for host function

TE activity does not always have to be harmful, as recent studies have begun to demonstrate. TEs contribute to genome evolution through TE exaptation, a process whereby TEs, which usually persist by replicating in the genome, transform into novel host genes or endogenous regulatory sites, which persist by conferring phenotypic benefits. In the human genome, important genes such as Rag1 & Rag2 [31], Syncytin [32], and Arc [33] have exapted from TEs.

In addition to providing novel proteins, TEs have been found as a species-specific transcription factor binding sites [34–36] acting as regulatory sequences [37–40] found to affect gene expression in different contexts. More specifically, [41] found an interferon inducible enhancer affecting gene expression of an interferon induced gene, AIM2, revealing TEs involvement in the regulation of essential immune functions.

TE exaptation in disease

TEs have also been found to be exapted to drive oncogene expression in human cancers. This process, termed onco-exaptation, involves the utilization of TE-derived promoters or enhancers, normally dormant or repressed, to aberrantly activate genes that contribute to cancer development. Three independent studies have found de-repression of TEs in B lymphocyte

immune cells through epigenetic alterations uncovering transcription factor binding sites driving expression of genes specific to Lymphoma cancer cells - FABP7, CSF1R, and IRF5 [42–44].

Further, recent work in the Chuong lab has found that oncogenic AP1/MAPK signaling drives the activation of enhancers derived from the primate-specific TEs which regulate tumor-specific expression of multiple genes associated with tumorigenesis, such as ATG12 and XRCC4 [45].

TEs play a dual role in the genome driving normal gene expression, but also contributing to the development of diseases when their activity is mis-regulated. Understanding the mechanisms of TE exaptation and their impacts on health and disease provides valuable insights into the complexity of the genome and offers potential targets for therapeutic intervention.

Using the idea that disease specific de-repression of TEs can provide promoters, enhancers, and other regulatory elements that influence the disease specific expression of nearby gene, I wanted to explore TEs in an aging and inflammatory context looking at TEs in oncogene-induced senescent cells.

Inflammation/Aging

Cellular senescence is a state of stable and long-term loss of proliferative capacity in cells, which occurs in response to various stressors. This process was first described by Hayflick and Moorhead in 1961, when they observed that human fibroblasts have a limited capacity to divide in culture, eventually entering a state of irreversible growth arrest despite remaining metabolically active [46]. As organisms age, senescent cells accumulate in various tissues. These cells do not function properly and can disrupt tissue architecture and function by secreting a variety of pro-inflammatory cytokines, chemokines, growth factors, and proteases known as

Senescence-Associated Secretory Phenotype (SASP) [47]. The SASP leads to chronic, low-grade inflammation, known as "inflammaging," which is a key driver of many age-related diseases [48–50].

De Cecco et al [51] sought out to understand the mechanism behind SASP. During cellular senescence, L1 elements (TE elements) become transcriptionally derepressed. This de-repression leads to the accumulation of L1-derived DNA in the cytoplasm of senescent cells. The presence of L1-derived cytoplasmic DNA activates a type-I interferon (IFN-I) response. This response is a hallmark of late senescence and contributes to the maintenance of SASP. Treatment of aged mice with lamivudine, a reverse transcriptase inhibitor, was shown to downregulate IFN-I activation and reduce age-associated inflammation in multiple tissues. Their finding suggests that targeting L1 reverse transcriptase could be a viable strategy for mitigating inflammaging and its related pathologies.

Research direction

TEs have been explored in the context of forming viral-like transcripts and therefore stimulating an interferon response [51] for its contribution to SASP, however, TEs have not been explored for its gene regulatory function in oncogene-induced senescent cells. I sought out to explore if there are any de-repressed oncogene-induced senescent specific TEs that are affecting gene expression associated to disease - SASP, immune dysregulation, or epigenetic modifiers. I found several disease related genes that were enhancer acting distance of oncogene-induced senescent specific TEs (see senescence appendix section), but only focused on one, SP140, for this Thesis.

In chapter II, I will discuss how a TE-derived isoform of SP140 is specific to various non-immune cancers, deviating from the oncogene-induced senescence model. SP140 was an attractive candidate due to its link to regulating inflammation [52–54]. However, cancer became the focus with SP140 due to oncogene-induced senescence being a type of cancer model as well. This model induced IMR90 lung fibroblast cells with H-RAS [55], making this a cell model of senescence and cancer. Further work can be done looking at other genes (Figure 8) that could drive the focus back on senescence and SASP.

Chapter II: An endogenous retrovirus regulates tumor-specific expression of the immune transcriptional regulator SP140

This chapter was adapted and expanded from published work[56]

Introduction

Chromatin readers facilitate transcriptional regulation by affecting chromatin structure binding or recruiting nuclear-signaling machinery at post-translational modification sites [57,58]. *Speckled Protein 140* (*SP140*) is a chromatin reader that is increasingly appreciated to influence immunological processes, inflammatory regulation, and disease [59], and is an interferon-stimulated gene predominantly expressed in immune cells including mature B cells, plasma cells, and some T cells [60–62]. *SP140* regulates transcriptional programs in human leukocytes where it represses topoisomerase and lineage-inappropriate genes and is indispensable for intestinal and immune homeostasis [63,64]. In lymphoblastoid cells, *SP140* represses the nuclear factor kappa-B (NF-kB) inflammatory response [54,65]. In monocyte-derived dendritic cells, *SP140* regulates maturation and tolerogenic potential [66]. In mice, *Sp140* acts as a negative regulator of type I interferon (IFN) responses that are essential for protection against intracellular bacterial infections, particularly *Mycobacterium tuberculosis* [53,67].

SP140 belongs to the speckled protein (SP) family of readers including *SP140*, *SP100*, *SP110*, and *SP140L*. All SPs contain 3 'reader' domains: the SAND domain named after Sp100, Aire, NucP41/P75, and Deaf-1, plant homeodomain (PHD), and bromodomain (BRD), a nuclear localization signal (NLS), and a caspase activation and recruitment domain (CARD). The SAND domain binds DNA directly or directs protein-protein interactions [68], and the PHD domain reads

histone methylation [69] with SP140 PHD also promoting intramolecular SUMOylation of the adjacent BRD domain [70]. The BRD binds acetylated histones, with SP140 BRD also binding promiscuously to acetylated H3 or H4 histone peptides [71]. Lastly, the CARD domain is important for hetero/homo-dimerization, allowing large protein complexes to form [72]. Uniquely, primate SP140 contains an intrinsically disordered region (IDR) allowing liquid-liquid phase separation, which has been attributed to mediate protein-protein interactions, signal transduction in immune cells, and phase separation of transcriptional machinery important for gene regulation [73–76].

Genome-wide association studies (GWAS) have unveiled that dysfunction of *SP140* is linked to several diseases, including Crohn's disease [77,78], multiple sclerosis [79], and chronic lymphocytic leukemia [80]. These diseases are associated with single nucleotide polymorphisms (SNPs) found within the introns of *SP140*. These non-coding SNPs alter *SP140* mRNA splicing leading to an overall reduction in the expression of the canonical *SP140* mRNA isoform [81], and more specifically, leading to elevated expression of *SP140* isoforms devoid of one or two exons part of the IDR (exon 7 for Crohn's and exons 7 and 11 for multiple sclerosis) [64,82].

In addition to playing a role in immunological diseases, there is emerging evidence that *SP140* expression plays a role in cancer pathogenesis. In both pan-cancer and head and neck cancer analysis, high expression of canonical *SP140* in tumor-associated macrophages negatively regulates transcription and phosphorylation of STAT1 and induces interferon-gamma signaling leading to higher tumor mutation burden, improved patient survival, and a favorable response to immunotherapy [83]**.** Further, a recent study focused on correlating epigenetic factors with the osteosarcoma cancer immunity cycle–a series of stepwise events required for anti-cancer immunity [84]. They identified *SP140* as an epigenetic protective factor highly correlated with prognosis, suggesting *SP140* to be a viable target for immunotherapy in osteosarcomas [85].

Outside of these studies, however, the potential role of SP140 in cancer remains largely unexplored.

Here, we investigated SP140 expression in cancer cells and found that *SP140* is aberrantly expressed in many human cancers of non-immune cell origin. Surprisingly, the tumor-specific expression of SP140 is driven by a non-canonical promoter, LTR8B, derived from a primate-specific endogenous retrovirus (ERV) located in the first intron. ERVs and other transposable elements are now recognized as influential players in gene expression regulation and splicing events, serving as enhancers, exons, and alternative promoters [86,87]. By analyzing transcriptomic data from patient cancer samples and cancer cell lines, we found recurrent expression of two novel SP140 isoforms driven by an LTR8B ERV that is predicted to generate both a truncated and nearly full-length SP140 protein. CRISPR perturbation of the LTR8B ERV-derived promoter in a human fibrosarcoma cell line, HT1080, showed that SP140 expression alters the response to IFN and influences the transcription of a few genes. Together, these findings implicate aberrant SP140 expression as a novel factor that may facilitate tumor immunity and transcriptional dysregulation.

Results

Cancer-specific expression of SP140 is driven by a novel LTR8B promoter

To investigate SP140 expression patterns, we first examined SP140 RNA expression across a panel of human cell lines and tissues profiled by the Human Protein Atlas. As expected, we found high expression of SP140 in immune cells, but surprisingly, we observed high

expression of the SP140 gene in many non-immune cell lines, including many cancer cell lines (Figure 2.1A). We inspected RNA-seq read alignments from these cells and discovered that SP140 expression in non-immune cells was exclusively driven by an intronic promoter derived from an LTR8B transposable element (Figure 2.2A). By conducting genome-guided transcriptome assembly on a subset of cell lines using StringTie2 [88] (Figure 2.1B), we confirmed that this LTR8B element drives the expression of two splice isoforms of SP140 (LTR8B-SP140): one previously annotated short isoform that is truncated after exon 3 (*LTR8B-SP140-Short*), and a novel near-full-length isoform that terminates at the canonical 3' end of the gene (*LTR8B-SP140-Long*) (Figure 2.2B). The *LTR8B-SP140-Short* isoform was previously annotated as a non-coding RNA, but potentially generates a truncated CARD domain. The novel *LTR8B-SP140-Long* isoform is predicted to generate a full-length SP140 protein except for the first 20 amino acids encoded within the canonical first exon (Figure 2.2B). Because these missing amino acids precede the characterized functional domains of SP140, the *LTR8B-SP140-Long* isoform may encode a protein that is functionally similar or identical to the canonical SP140 protein. Notably, there is high similarity of the first 20 amino acids of SP140 across primates, but this region is not conserved in mouse and other mammals, suggesting that this N-terminal region specific to the canonical isoform is not essential for SP140 function (Figure 2.1C-D). Given the established role for SP140 as a chromatin regulator of immune signaling in immune cells, these results suggest that aberrant expression of LTR8B-SP140 in cancer cells may play a role in cancer dysfunction.

We next asked whether the LTR8B-SP140 isoforms show evidence of expression in other RNA-seq databases. We examined junction counts using the recount2 database, which includes harmonized exon-exon junction counts across ~49 thousand RNA-seq datasets [89]. Using exon

junction information, we identified datasets containing spliced RNA-seq reads supporting the junction between the LTR8B promoter and the 2nd canonical exon of SP140 (LTR8B-Exon2). This analysis was unable to distinguish between the *LTR8B-SP140-Short* and *LTR8B-SP140-Long* isoforms. We identified robust expression of the LTR8B-Exon2 junction in 52 samples, most of which corresponded to tumors or cancer cell lines (Figure 2.2C). Further, looking at junction counts strictly from the Genome-Tissue Expression (GTEx) database, we were able to verify that the LTR8B-SP140 isoforms exhibit very low expression (LTR8B-Exon2 junction copies per million $< .25$) in normal non-cancerous tissues.

To specifically analyze cancer cell samples, we compared the LTR8B-Exon2 junction counts to the canonical SP140 Exon1-Exon2 junction counts in RNA-seq data from the Cancer Cell Line Encyclopedia (CCLE) representing ~1000 different cancer cell lines [90]. This analysis confirmed that LTR8B-SP140 isoforms are exclusively expressed in non-immune cell cancers, in contrast to the canonical SP140 isoform which is exclusively expressed in immune cell cancers (Figure 2.2D).

Finally, we expanded our analysis beyond cancer cell lines to patient tumors from The Cancer Genome Atlas. We used RJunBase to quantify junction read counts supporting the LTR8B-Exon2 junction, and found recurrent expression of the LTR8B-SP140 isoforms in a subset of patient tumors across multiple cancer types (Figure 2.2E) [91], in contrast to our GTEx analysis of normal tissues. Only a fraction (~1%) of all patient tumors showed expression of the LTR8B-SP140 isoform, indicating that only a subset of cancers expresses this isoform or that it is expressed in small cell populations within the tumor. Together, these transcriptomic analyses indicate that the LTR8B-SP140 isoforms are expressed in multiple non-immune cancer cell lines and patient tumors from multiple cancer types.

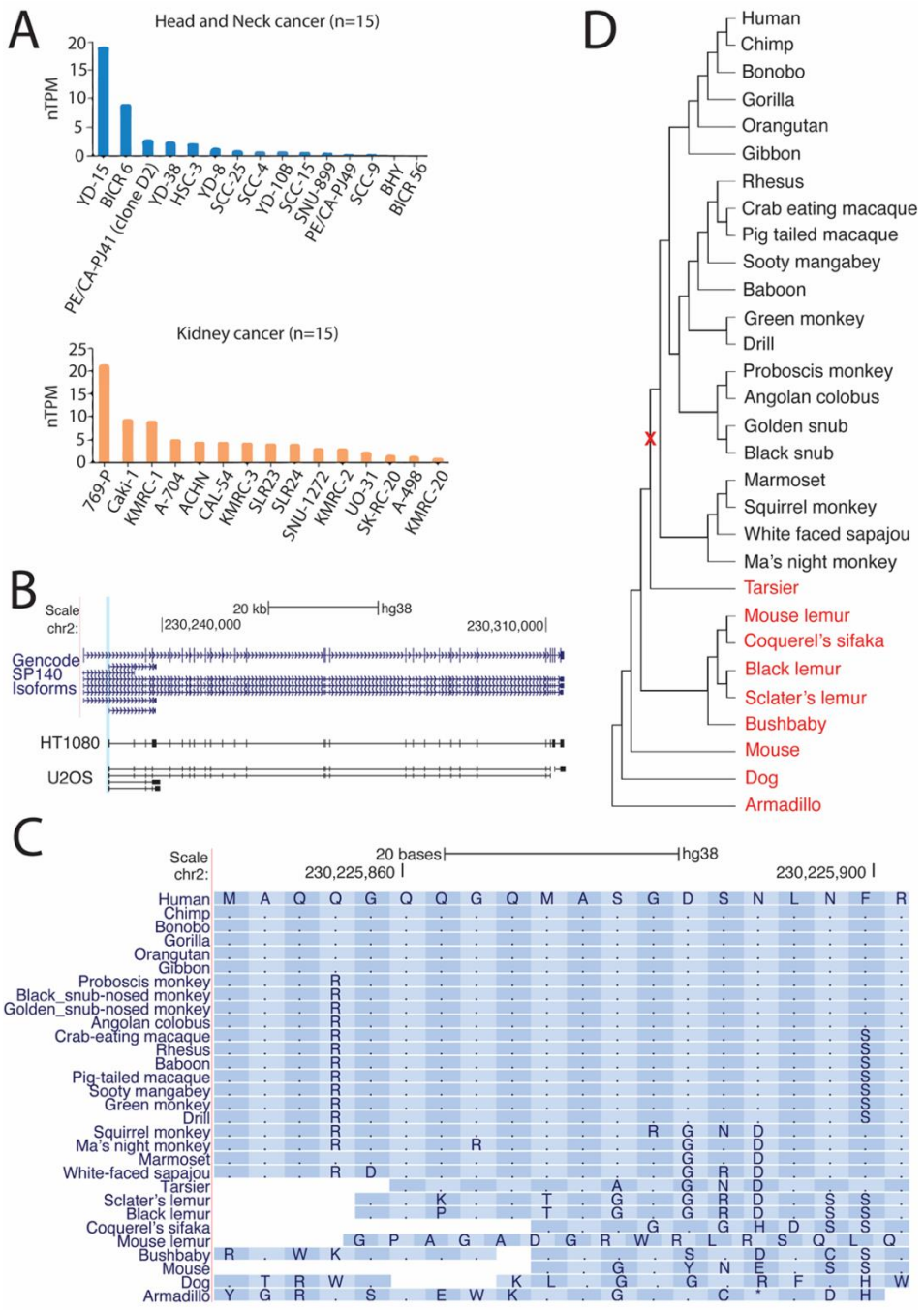


Figure 2.1. RNA-seq evidence for expression of SP140 in cancer cells and evolutionary analysis of first intron of SP140. **(A)** Screenshot of 2 cancers from the Human Protein Atlas of SP140 expression (https://www.proteinatlas.org/ENSG00000079263-SP140/cell+line), quantified at the gene level. Cell lines are on the x-axis, and the y-axis represents SP140 transcripts per million (TPM). **(B)** UCSC genome browser screenshot of HT1080 and U2OS StringTie2 transcriptome assembly at the SP140 locus. The blue highlight covers the LTR8B region within the first intron of SP140. **(C)** UCSC genome browser 'Cons 30 Primates Track' analysis of first 20 amino acids of SP140. Dots represent a conserved amino acid. **(D)** Phylogenetic tree of the 30 species from Figure 2.1C. The red 'X' represents point in the tree where the primate-specific first 20 amino acids emerge; species in red lack conservation to this region.

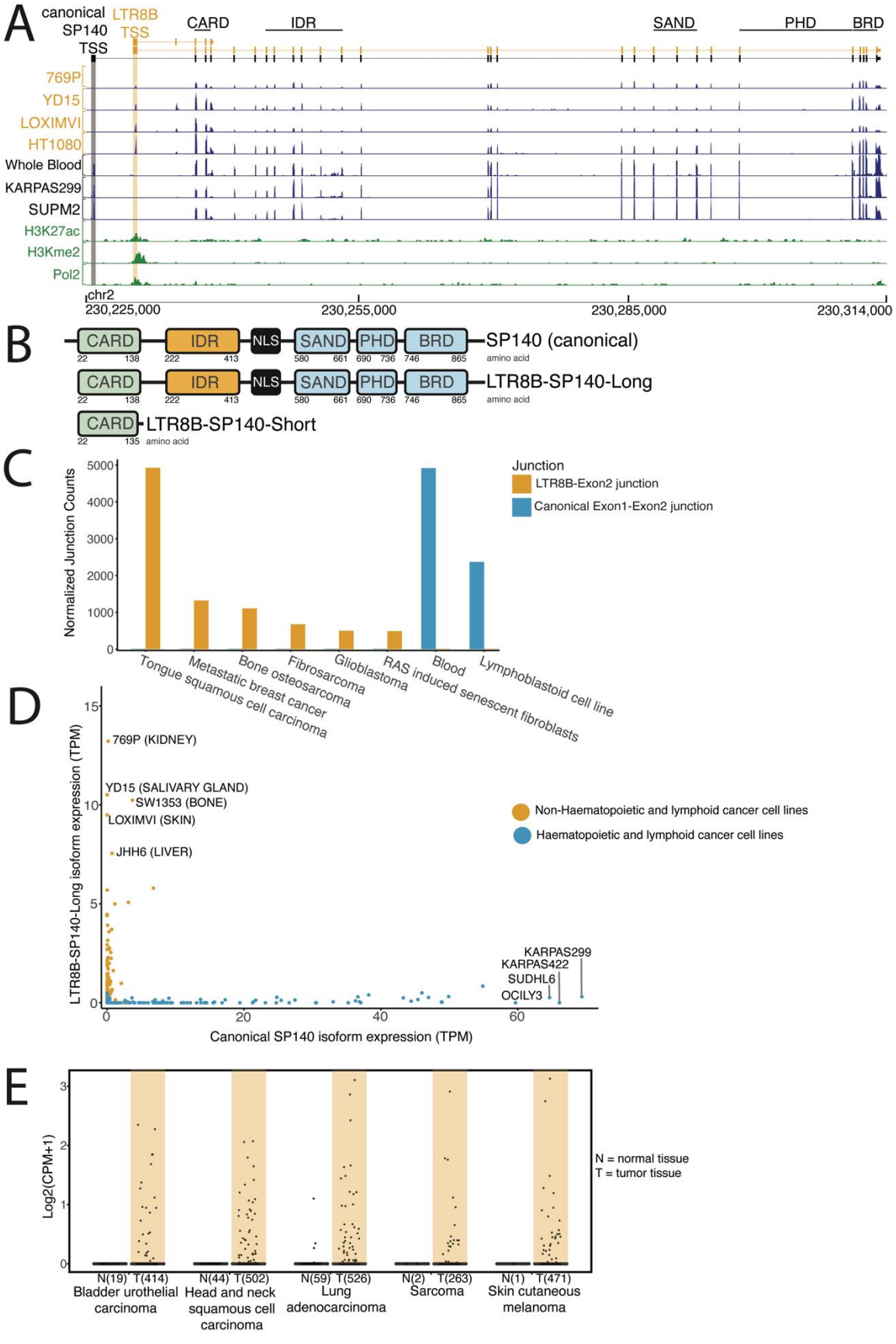

A
canonical SP140 TSS
LTR8B TSS
CARD
IDR
SAND
PHD
BRD
769P
YD15
LOXIMVI
HT1080
Whole Blood
KARPAS299
SUPM2
H3K27ac
H3Kme2
Pol2
chr2
230,225,000
230,255,000
230,285,000
230,314,000
B
CARD
IDR
NLS
SAND
PHD
BRD
SP140 (canonical)
amino acid
LTR8B-SP140-Long
LTR8B-SP140-Short
C
Normalized Junction Counts
Junction
LTR8B-Exon2 junction
Canonical Exon1-Exon2 junction
Tongue squamous cell carcinoma
Metastatic breast cancer
Bone osteosarcoma
Fibrosarcoma
Glioblastoma
RAS induced senescent fibroblasts
Blood
Lymphoblastoid cell line
D
LTR8B-SP140-Long isoform expression (TPM)
Canonical SP140 isoform expression (TPM)
769P (KIDNEY)
YD15 (SALIVARY GLAND)
SW1353 (BONE)
LOXIMVI (SKIN)
JHH6 (LIVER)
KARPAS299
KARPAS422
SUDHL6
OCILY3
Non-Haematopoietic and lymphoid cancer cell lines
Haematopoietic and lymphoid cancer cell lines
E
Log2(CPM+1)
N = normal tissue
T = tumor tissue
N(19) T(414)
Bladder urothelial carcinoma
N(44) T(502)
Head and neck squamous cell carcinoma
N(59) T(526)
Lung adenocarcinoma
N(2) T(263)
Sarcoma
N(1) T(471)
Skin cutaneous melanoma

Figure 2.2. An endogenous retrovirus drives cancer-specific expression of SP140.
(A) UCSC genome browser screenshot of RNA-seq and CUT&RUN of multiple cell lines at the *SP140* locus. The intronic ERV (LTR8B) alternative transcription start site (TSS) in various non-immune cancers produces two isoforms (orange). Cell lines in orange are RNA-seq from various non-immune cancer cell lines. Cell lines in black are RNA-seq from either a whole blood sample taken from GTEx or immune cancer cell lines showing canonical *SP140* transcription. Tracks in green are CUT&RUN in HT1080 cells. Each track is scaled at [0-6] TPM, except for whole blood [0-40] TPM. Domain locations are labeled at the top. Abbreviations: CARD = caspase activation and recruitment domain; IDR = intrinsically disordered region; SAND = Sp100, Aire, NucP41/P75, and Deaf-1; PHD = plant homeodomain; BRD = bromodomain. **(B)** *SP140* isoforms are broken down by functional domains with amino acid markers. The top isoform depicts a full-length canonical *SP140* isoform. The middle depicts a novel *LTR8B-SP140-Long* isoform missing only the first 20 amino acids and with all functional domains. The bottom depicts an *LTR8B-SP140-Short* isoform with only the CARD domain intact. Abbreviations described in (A). **(C)** SP140 junction counts in selected cell lines from recount2. Orange bars depict LTR8B-Exon2 junction counts of the *LTR8B-SP140* isoforms. Blue bars depict Exon1-Exon2 junction counts of the canonical *SP140* isoform. **(D)** Scatter plot of isoform quantification of Cancer Cell Line Encyclopedia (CCLE) RNA-seq data. Each dot represents a cancer cell line with blue dots depicting immune cancer cell lines (hematopoietic and lymphoid tissue) and orange dots depicting non-immune cancer cell lines. The X-axis depicts canonical *Sp140* isoform expression and the y-axis depicts the *LTR8B-SP140-Long* isoform transcripts per million (TPM). **(E)** LTR8B-Exon2 junction counts from 5 different human non-immune tumors, from The Cancer Genome Atlas. Each dot represents a patient. N = normal matched tissue; T = extracted tumor; (#) represents the number of patient samples. Tumor lines are highlighted in orange.

LTR8B elements show regulatory activation in HT1080 cancer cells

We next focused on HT1080 fibrosarcoma cells as an example cancer cell line that expresses the LTR8B-SP140 isoforms (Figure 2.2C). Given that transposable element-derived regulatory elements are often activated at a family level [41,45,92,93], we investigated whether LTR8B elements were globally activated in HT1080 cells. We used GIGGLE [94] to determine if LTR8B loci (N=1730) were colocalized with H3K27ac, a marker of regulatory elements. We found that LTR8B elements were significantly enriched within H3K27ac-marked elements (N=70; Fisher-two-tail $P = 2.9e^{-7}$), while we did not see enrichment for the related subfamilies LTR8 and LTR8A (Figure 2.3A-B). This suggests that LTR8B elements contain sequence motifs that facilitate their regulatory activation in cancer cells.

LTR8B elements were recently reported to show placenta-specific enhancer activity, where they are regulated by the JUND transcription factor of the AP1 complex [95–98]. To investigate how LTR8B is regulated in HT1080 cells, we conducted motif enrichment analysis of the 70 H3K27ac-marked LTR8B loci, using the remaining 1,660 LTR8B loci sequence as the background set [99]. This analysis identified the AP1 components FOSL2::JUN, JUNB, and FOSB::JUN motifs as hits within H3K27ac-marked LTR8B loci. Finally, we used the Cistrome database [100] to globally search for transcription factors and chromatin regulators that are significantly correlated with LTR8B elements. We found that the 70 H3K27ac-marked LTR8B loci were enriched for multiple AP1 motifs (JUND, FOS, JUN, FOSL1, FOSL2) (Figure 2.4A), while our 1,660 LTR8B loci background set (Figure 2.4B) were not enriched for any AP1 motifs. Our analysis indicates that in addition to their previously reported placenta-specific activity [95–98], LTR8B elements also show regulatory activity in HT1080 fibrosarcoma lines and other cancer

cell types, likely due to shared activation of mitogen-activated protein kinase (MAPK) and activator protein 1 (AP1) signaling.

The AP1 signaling pathway is frequently dysregulated in cancers with mutations in the MAPK signaling pathway [101]. To test whether MAPK/AP1 signaling is required for LTR8B-SP140 expression, we treated HT1080 cells with the MAPK/AP1 activator Tumor Necrosis Factor-alpha (TNF-α) and measured LTR8B-SP140 (*LTR8B-SP140-Long* & *LTR8B-SP140-Short)* expression by qRT-PCR. We found that LTR8B-SP140 expression was increased upon MAPK/AP1 activation (Figure 2.3C), with no effect of canonical SP140 (Figure 2.4C), demonstrating that MAPK/AP1 signaling drives aberrant expression of the LTR8B-SP140 isoform.

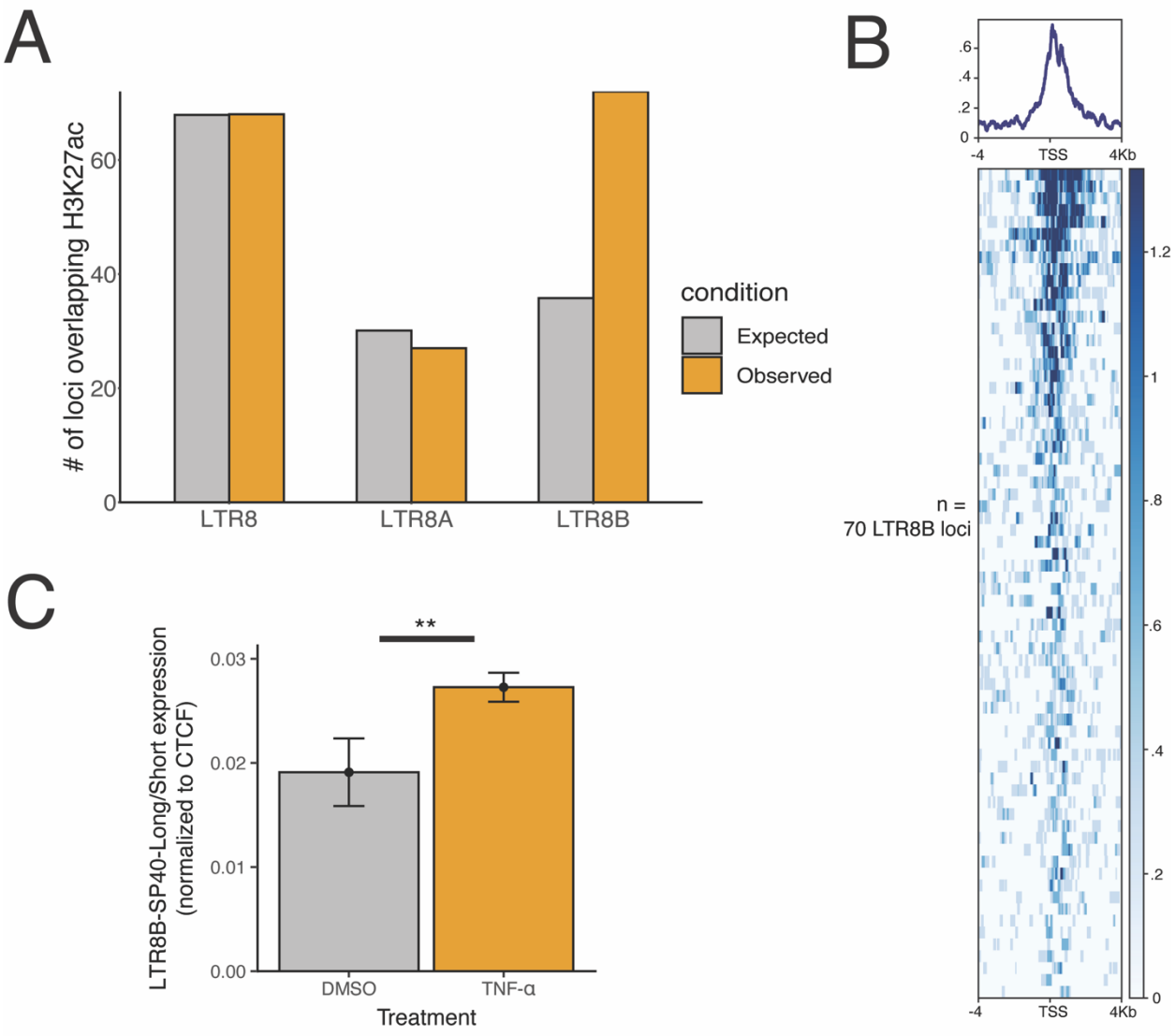


Figure 2.3. The endogenous retrovirus subfamily LTR8B is de-repressed in HT1080 cells. **(A)** Giggle results of the LTR8/A/B family for the number of expected CUT&RUN H3K27ac overlaps vs observed, with the respective ERV (LTR8, LTR8A, LTR8B). Fisher's two tail test for LTR8, LTR8A, and LTR8B are 1.0, $5.3e^{-1}$, and $2.8e^{-7}$, respectively. **(B)** H3K27ac CUT&RUN peak heatmap of the 70 LTR8B loci. **(C)** RT-qPCR results of LTR8B-SP140 expression (*LTR8B-SP140-Long* & *LTR8B-SP140-Short)* in HT1080 cells treated with DMSO (left; grey) or treated with TNF-α (right; orange) for 24 hours. 6 replicates are used. A paired t-test was used for significance: *(***)* p-value < .01. Error bars represent the minimum and maximum values of data.

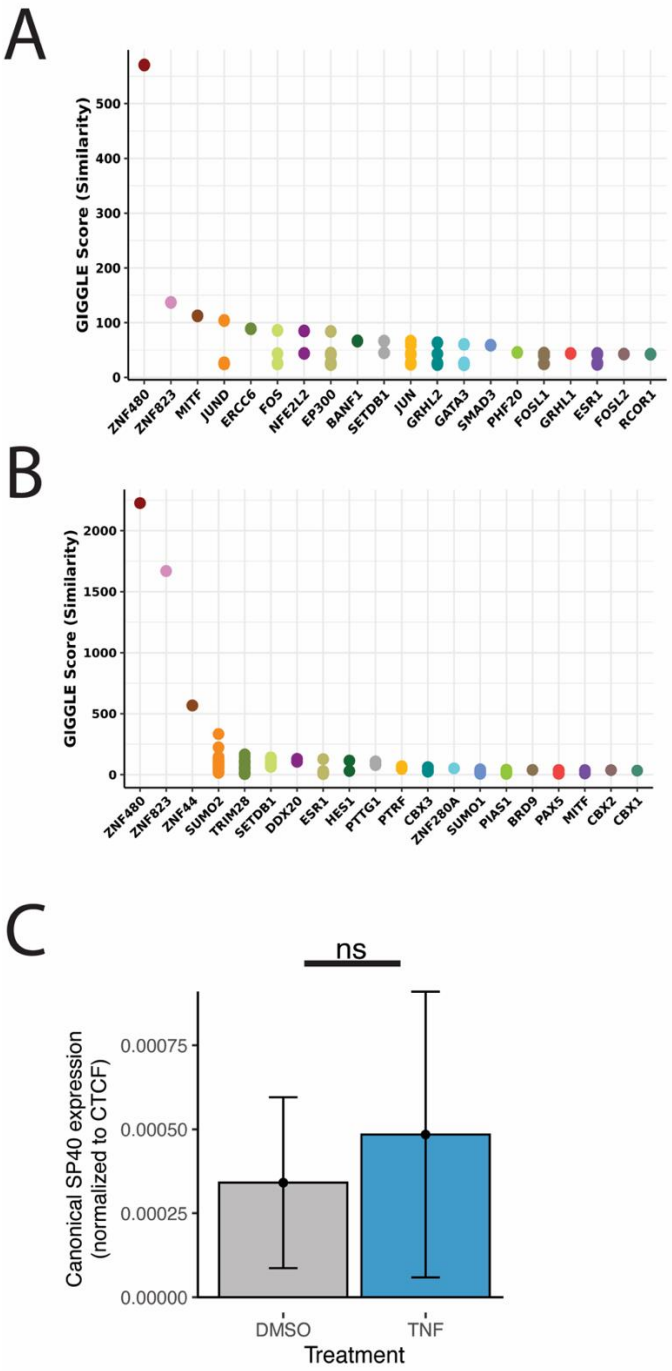


Figure 2.4. Cistrome analysis of transcription factors colocalized with active or inactive LTR8B elements and canonical SP140 expression in HT1080s. **(A)** Cistrome output of top transcription factors and chromatin regulators that colocalize with the 70 activated LTR8B loci which overlap with H3K27ac in HT1080 cells. **(B)** Cistrome output of top transcription factors and chromatin regulators that colocalize with the 1,660 background LTR8B loci which do not overlap with H3K27ac in HT1080 cells. **(C)** RT-qPCR results of Canonical SP140 expression in HT1080 cells treated with DMSO (left; grey) or treated with TNF-α (right; blue) for 24 hours. 6 replicates are used. A paired t-test was used for significance: (ns) p-value >.05. Error bars represent the minimum and maximum values of data. 6 replicates are used.

The LTR8B-SP140-Long isoform encodes a nuclear protein

We next determined whether the *LTR8B-SP140-Long* isoform encoded a protein, similar to the canonical SP140. While the *LTR8B-SP140-Short* isoform may be functionally significant in cancer cells, we chose to focus on the *LTR8B-SP140-Long* isoform, given its predicted similarity to the canonical SP140 protein, which is localized in nuclear speckles and has been functionally characterized as a chromatin regulator in immune cells [52,64]. To validate that the long *LTR8B-SP140-Long* isoform encodes a protein, we used CRISPR to add a HiBiT epitope tag to the canonical 3' end of SP140. We tagged *SP140* to ensure that we were not detecting *SP140L*, a highly similar paralog of SP140 [102] which is also expressed in HT1080 cells. After isolating a clonal line with a homozygous HiBiT insertion, we confirmed by luminescence that the HiBiT tag was expressed (Figure 2.5). Using an antibody against the HiBiT epitope, we confirmed by Western blotting that the tagged LTR8B-SP140-Long protein comes out to be ~80kDa (Figure 2.6A). Finally, by immunofluorescence imaging, we determined that *LTR8B-SP140-Long* is localized to nuclear speckles (Figure 2.6B). Together, these results suggest that the *LTR8B-SP140-Long* isoform forms a stable protein that has similar nuclear localization as the canonical *SP140* isoform, which should be further validated co-staining with a speckles marker, supporting the idea that the LTR8B-SP140-Long isoform may have similar functional activity in cancer cells.

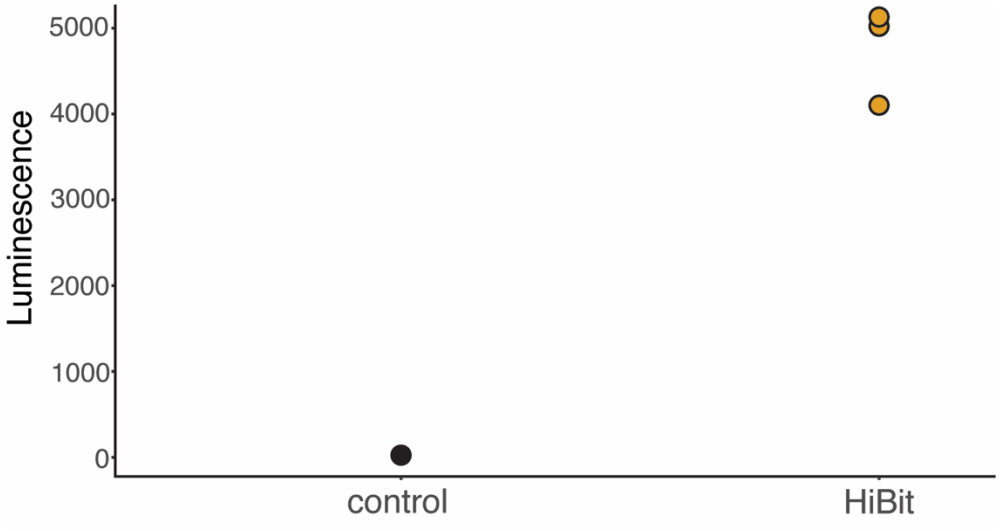


Figure 2.5. Supporting evidence for LTR8B-SP140-Long protein expression.
Box plot of luminescence output from HiBit Lytic detection assay of control HT1080 cell line and a c-terminal HiBiT SP140 tagged cell line. 3 replicates are used.

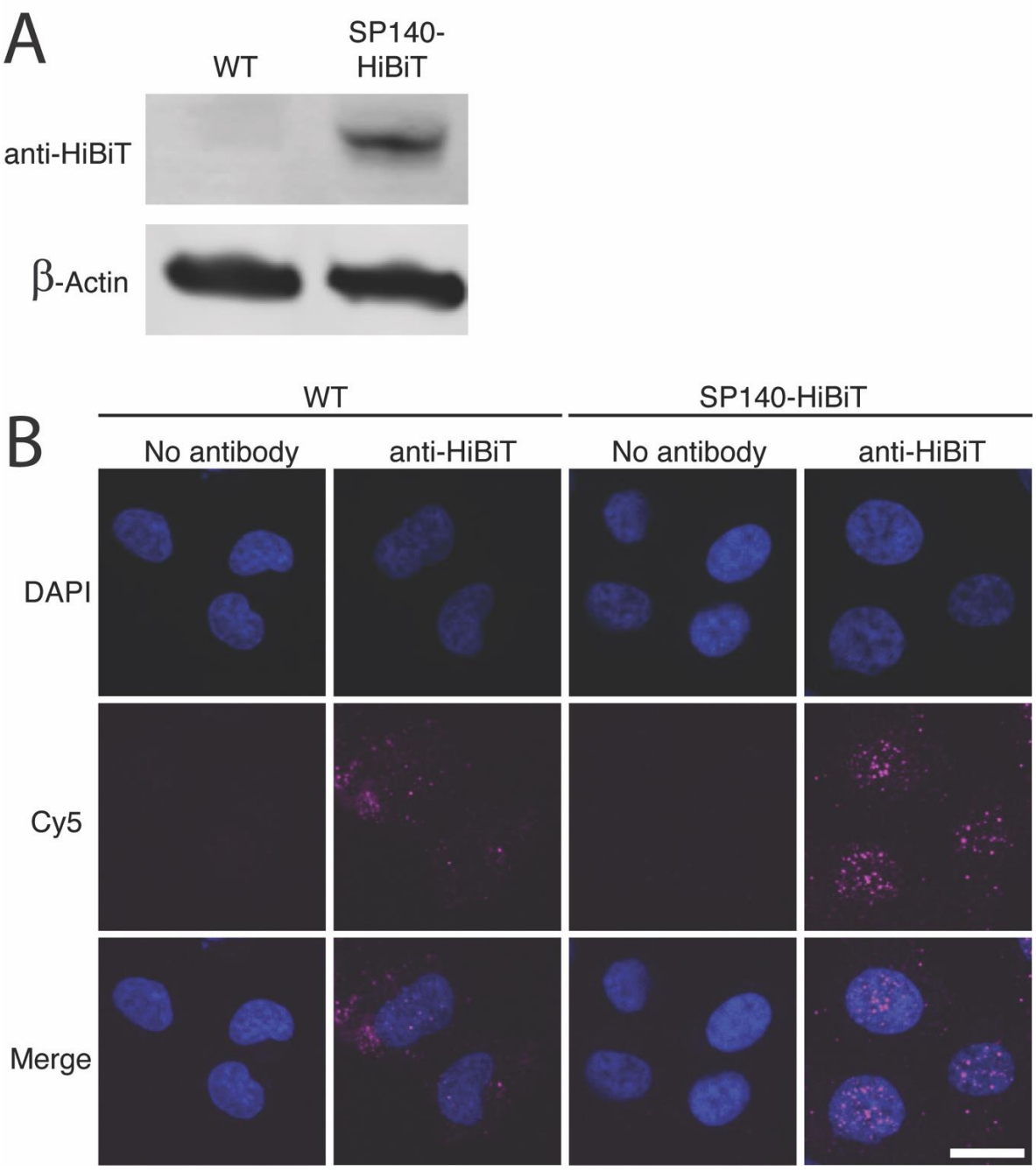


Figure 2.6. The LTR8B-SP140-Long isoform encodes a protein with similar localization as the canonical SP140 isoform. **(A)** Western blot showing expression of endogenously tagged SP140-HiBiT in HT1080 cells, using the anti-HiBiT antibody. B-Actin was used as a loading control. **(B)** Immunofluorescence of HT1080 wild-type (WT) and SP140-HiBiT tagged cells. The DAPI panel shows nucleus staining and Cy5 panel shows anti-HiBiT antibody labeling. Scale bar = 20um.

Silencing the LTR8B promoter of SP140 inhibits interferon mediated cytotoxicity in cancer cells

To investigate the potential functional impact of the LTR8B-driven expression of SP140, we conducted CRISPRi to silence the isoforms by targeting the LTR8B-SP140 element. We generated HT1080 cells stably expressing the dCas9-KRAB-MeCP2 construct [103] and transfected cells, either with non-targeting negative control guide RNA (gRNA) or a gRNA specific to the LTR8B-SP140 element with no off-target binding predicted to other LTR8B elements. We note that silencing this promoter is expected to silence both short and long isoforms driven by this promoter. We confirmed 88% knockdown of both the long and short isoforms silencing through qRT-PCR (Figure 2.7) and RNA-seq. Furthermore, we found that silencing the LTR8B element did not affect any nearby genes, including *SP140L,* with the nearest affected gene, INPP5D, being ~3 megabases away (Figure 2.8A).

Given previous reports that canonical *SP140* negatively regulates immune signaling in immune cells [64], we tested how silencing LTR8B-SP140 affects the response to type I IFN by assaying cytotoxicity after extended IFNβ treatment. After 3 days of IFNβ treatments at various concentrations, our LTR8B-SP140 silenced cells showed decreased cell viability compared to the control cells (Figure 2.8B). The decrease of viability in cells where LTR8B-SP140 was silenced indicates that *LTR8B-SP140-Long* and/or *LTR8B-SP140-Short* have immunomodulatory activity, similar to the canonical isoform function in immune cells [54,63–66].

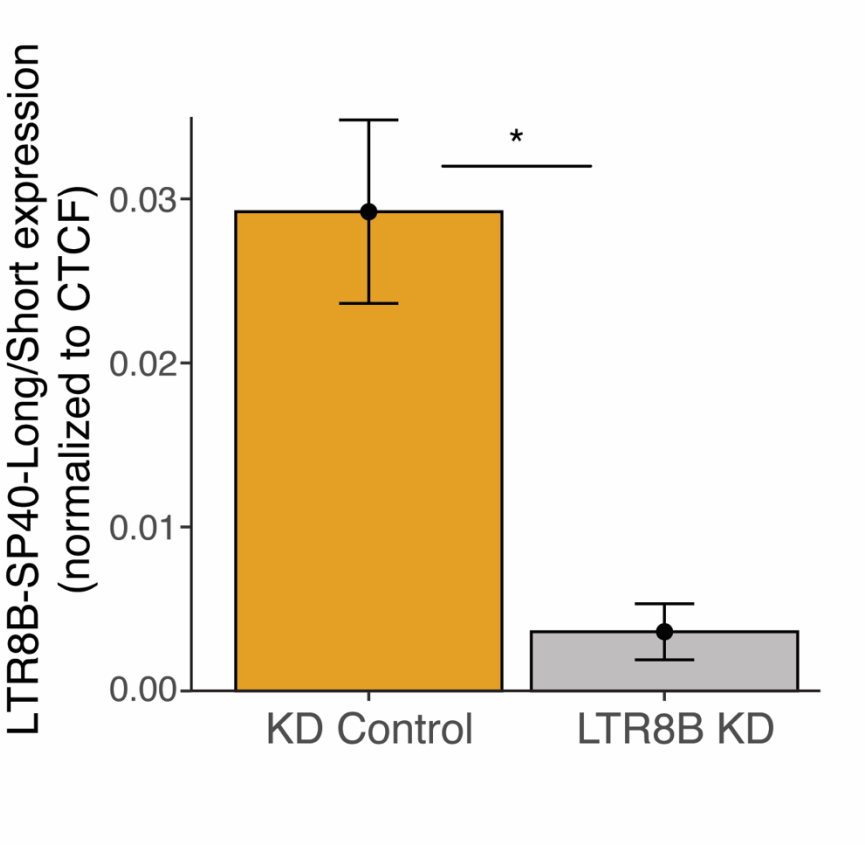


Figure 2.7. Supporting evidence for LTR8B-SP140 isoform repression.
RT-qPCR results of LTR8B-SP140 expression (*LTR8B-SP140-Long* & *LTR8B-SP140-Short)* in HT1080 cells with a CRISPRi control guide (left; orange) or a CRISPRi guide targeting the intronic SP140 LTR8B region (right; grey). Error bars represent the minimum and maximum values of data. A two-sample equal variance t-test was used for significance:(*) p-value < .05. 6 replicates are used.

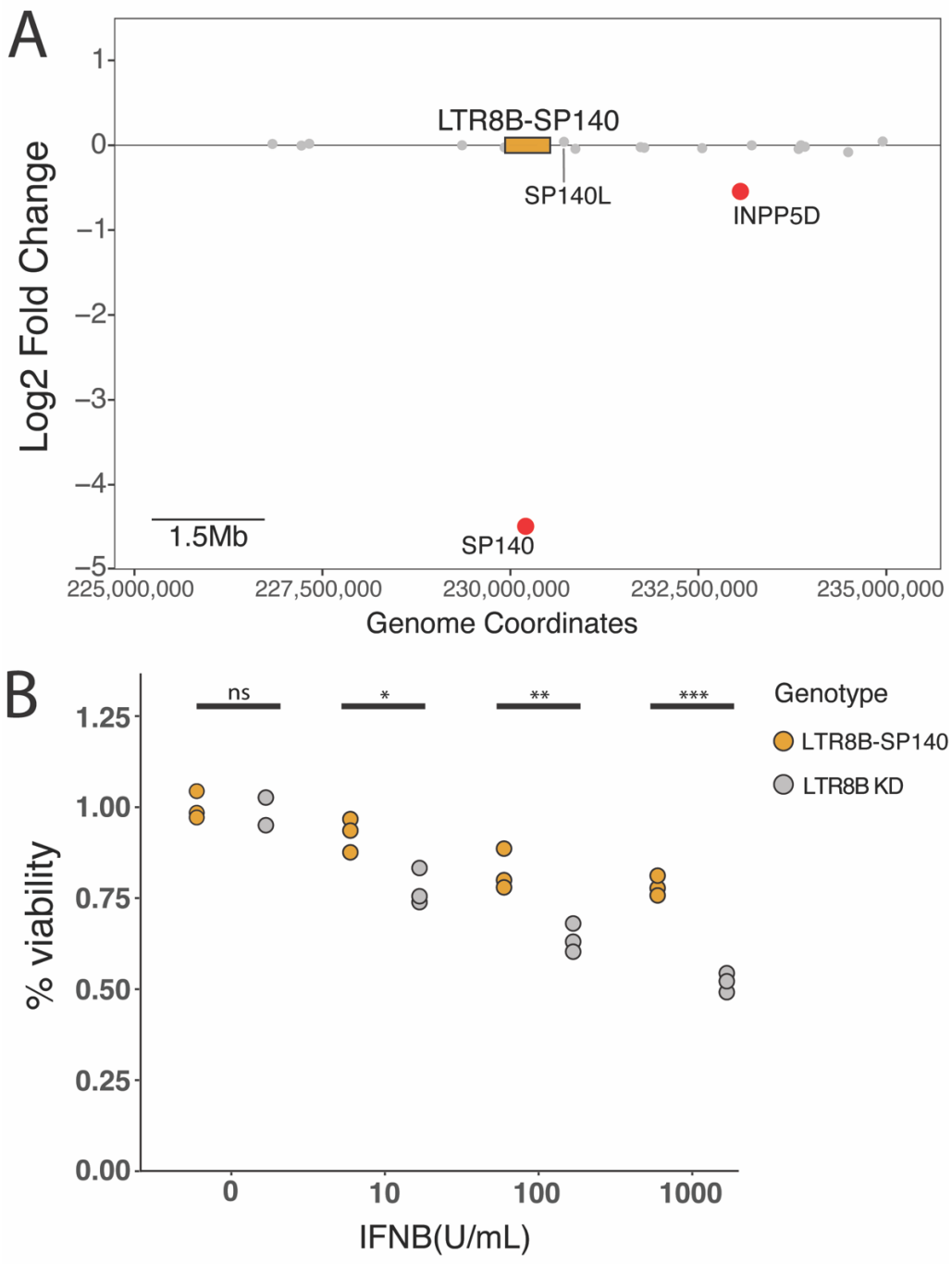


Figure 2.8. Silencing the LTR8B promoter of SP140 causes a stronger cytotoxic response to IFNβ in cancer cells. **(A)** Distance plot of differentially expressed genes of SP140 KD RNA-seq centered at the *LTR8B-SP140-Long* locus (orange box; not to scale). **(B)** Cell viability of wild-type LTR8B-SP140 HT1080 (orange) and LTR8B-SP140 KD HT1080 (grey) cells treated with increasing concentrations of IFNβ (0, 10, 100, & 1000 U). 3 replicates are used. A two-sample equal variance t-test was used for significance:(*) p-value < .05; (**) p-value < .01; (***) p-value < .001.

Silencing the LTR8B promoter of SP140 leads to gene derepression

Given our finding that the SP140 protein isoform encoded by LTR8B-SP140-Long localizes to nuclear speckles similarly to the canonical full-length SP140 isoform (Figure 2.6B), we asked how LTR8B-SP140 expression would affect gene regulation in non-immune cells. We conducted RNA-seq in HT1080 cells where LTR8B-SP140 isoforms were silenced by CRISPRi and compared it to HT1080 cells transfected with a negative CRISPRi control. Our differential expression analysis showed that silencing LTR8B-SP140 resulted in significant upregulation of 19 genes (padj< 5e-2 & log2FoldChange > 1) and downregulation of 10 genes including SP140 (padj< 5e-2 & log2FoldChange < -1) (Figure 2.9). These results are consistent with the established activity of canonical SP140 as a transcriptional repressor. However, we did not observe upregulation of the same genes previously reported in human macrophages, including developmental HOX transcription factors [64]. Notably, the most significantly upregulated gene, Leucine-Rich Repeats And Calponin Homology Domain Containing 3 (LRCH3), is important in activating the transcription factor NFkB, which is a pivotal mediator of immune response, inflammation, cell growth/survival, and development, where aberrant expression contributes to autoimmunity, inflammatory diseases, and malignant disorders [104,105]. Suppression of this gene may contribute to the immunomodulatory function of *LTR8B-SP140-Long* that we previously observed (Figure 2.8B). To determine if these differentially expressed genes (DEGs) were a result of off-target binding of our CRISPRi LTR8B-SP140 target, we checked if our DEGs occurred within common enhancer acting distance (50kb) of LTR8, LTR8A, and LTR8B. We found that SP140 was the only downregulated gene within 50kb of all LTR8s suggesting that our LTR8B-SP140 target is specific (Figure 2.10A), however, implementing CRISPRi guides other than the LTR8B-SP140 locus to disrupt SP140 function would further conclude no off-target

effects. We also profiled expression after IFNβ treatment and found similar differential gene profiles of SP140 KD vs control with or without IFNβ treatment (Figure 2.10B). Unlike the canonical SP140 isoform [62], neither LTR8B-SP140 isoform was induced by IFNβ, consistent with their usage of a different promoter. Altogether, these findings are consistent with the activity of *LTR8B-SP140-Long* as a transcriptional repressor in cancer cells, with the caveat that some these effects may also be caused by the *LTR8B-SP140-Short* isoform, which is also silenced in these experiments.

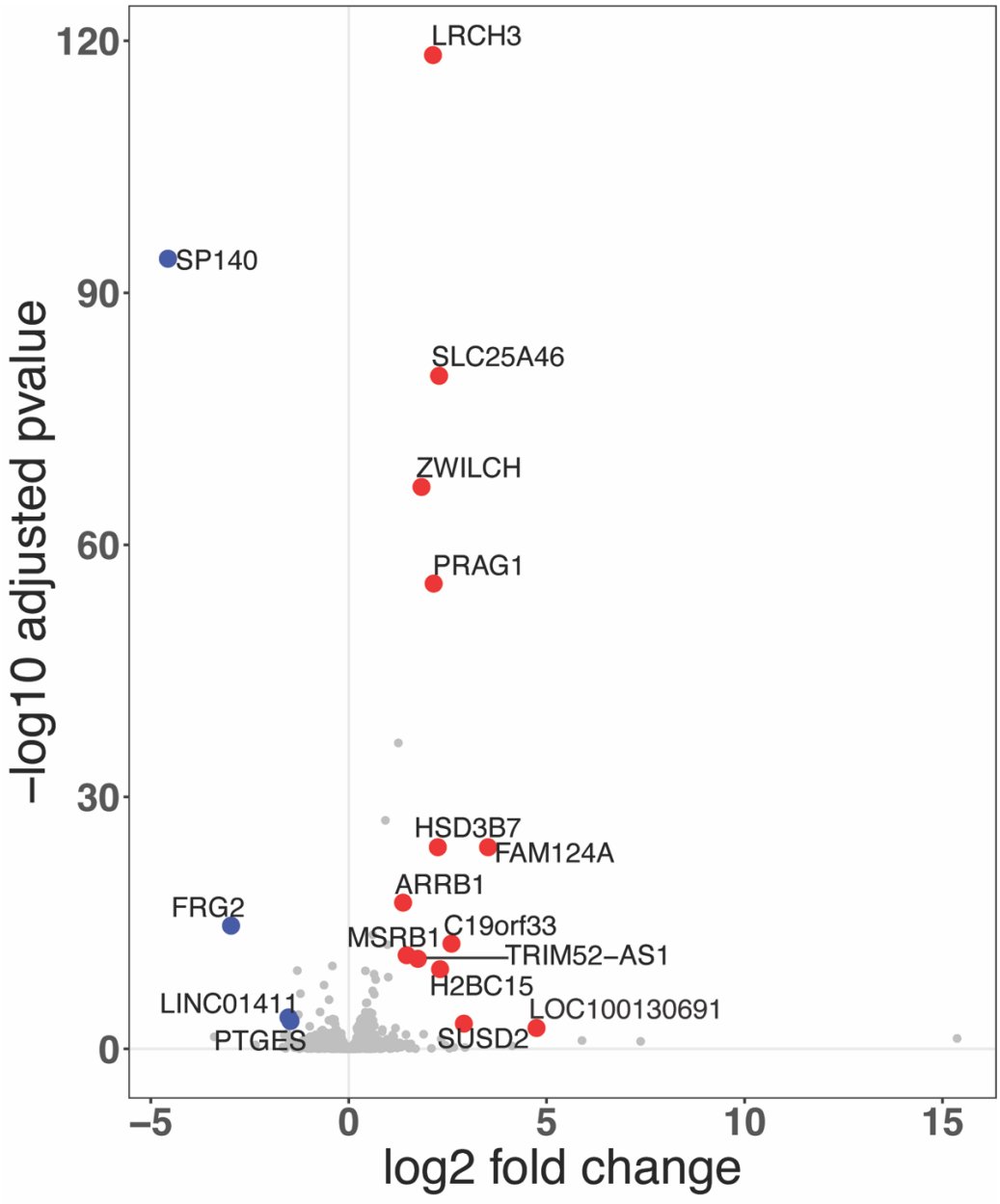


Figure 2.9. Silencing the LTR8B promoter of SP140 causes gene dysregulation. Volcano plot of differentially expressed genes using DeSeq2 between LTR8B-SP140 KD HT1080 vs dCas9-HT1080 cell lines. Each dot represents a gene. Positively expressed genes upon SP140 KD are in red and negatively expressed genes upon SP140 KD are in blue. Label cutoffs at 5e-3 for adjusted p-value and +/- 1.3 log2foldchange. 2 replicates are used.

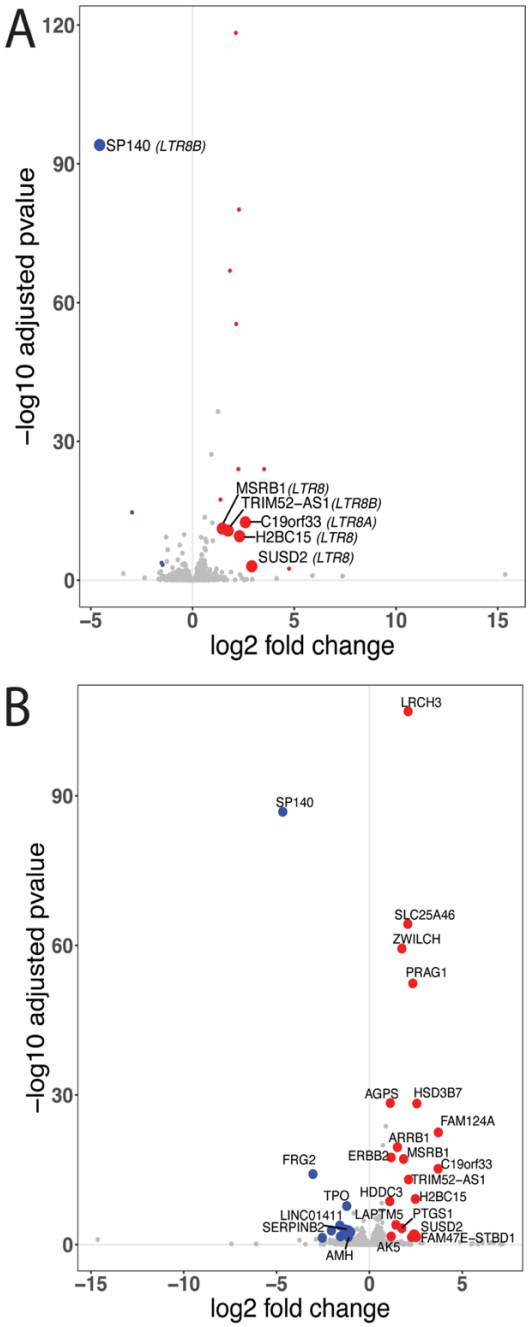


Figure 2.10. Differentially expressed genes within 50kb of LTR8s and RNA-seq differential expression of LTR8B-silenced cells in IFNβ-stimulated conditions. **(A)** Volcano plot of differentially expressed genes using DeSeq2 between LTR8B-SP140 KD HT1080 vs dCas9-HT1080 cell lines. Each dot represents a gene. Positively expressed genes upon SP140 KD are in red and negatively expressed genes upon SP140 KD are in blue. This is same as Figure 2.9, except labeled genes are within 50kb of LTR8, LTR8A, and LTR8B. 2 replicates are used. **(B)** Volcano plot of differentially expressed genes using DESeq2 between LTR8B-Sp140 KD HT1080 cells treated with 10U/mL IFNβ for 4 hours vs dCas9-HT1080 cells treated with 10U/mL IFNβ for 4 hours. Each dot represents a gene. Positively expressed genes upon SP140 KD are in red and negatively expressed genes upon SP140 KD are in blue. Label cutoffs at 5e-2 for adjusted p-value and +/- 1 log2foldchange. 2 replicates are used.

Discussion

Our study shows that the immune chromatin regulator *SP140* is aberrantly expressed in many cancer types, driven by an ERV-derived promoter located in the first intron. Notably, although the LTR8B element functions as an alternative promoter, the insertion is antisense to the gene, suggesting non-canonical promoter activity. While previous studies have identified *SP140* expression in tumors, the expression of *SP140* was attributed to tumor-associated macrophages [83]. Our RNA-seq analysis revealed that aberrant expression of *SP140* in cells from multiple non-immune cancers results in transcription of both a short truncated isoform and a nearly full-length isoform predicted to encode a protein with 99% similarity to the canonical SP140 protein. Using HT1080 fibrosarcoma cells as a model, we found that the aberrant expression of *SP140* inhibits IFN-mediated cytotoxicity and causes an altered transcriptional landscape. Notably, the altered transcriptional landscape does not match the canonical SP140 repressive patterns reported in macrophages [64]. This may be because SP140 function evolved in the context of immune cells, making their activity when reactivated in cancer cells largely random and influenced by the epigenetic landscape of the cell. Our findings implicate aberrant *SP140* expression, driven by an intronic LTR8B-derived promoter, as a contributing factor to the pathogenesis of a broad range of cancers.

Our understanding of the biological function of *SP140* remains incomplete. *SP140* is a chromatin regulator that associates with the Polycomb complex and suppresses lineage-inappropriate cells in immune cells [63,64]. *SP140* also suppresses innate immune responses including the type I IFN response, yet how *SP140* mechanistically plays a role in these processes remains unclear. Our finding that *SP140* is also expressed in cancer cells provides an opportunity

to dissect *SP140* function in non-immune cell types and determine which aspects of *SP140* function are cell type specific.

Our finding adds to our understanding of how alternative or aberrant splicing of SP140 can contribute to disease. The canonical form of *SP140* is expressed in immune cells, but several studies have identified splicing defects in *SP140* as an underlying cause of diseases like Crohn's disease [77,78]. Our study reveals widespread expression of a novel splice isoform of *SP140* and a previously annotated short isoform predicted to not encode a protein, expressed specifically in cancer cells, which may contribute to cancer cell epigenomic dysregulation and pathogenesis.

There are several limitations to our study. Our functional studies support a role for SP140 in altering the immune phenotype and transcriptional landscape of cancer cells, but further studies will be needed to fully define the consequences of SP140 on cancer growth or therapy resistance, particularly *in vivo*. Although our TCGA patient tumor analysis showed low occurrence (~1%) of the LTR8B-SP140 isoform, this could be due to cell heterogeneity, which single-cell RNA-seq analysis would be important to define the expression of the isoform. Further, the SNP-associated splice variants of Crohn's disease, multiple sclerosis, and chronic lymphocytic leukemia [77–80] would be interesting to examine amongst the non-immune cell cancers expressing LTR8B-SP140 to see if there are any differences in isoform expression. Additionally, it would be interesting to screen for the LTR8B-SP140 isoforms in patients with the SNP-associated splice variants of Crohn's disease, multiple sclerosis, and chronic lymphocytic leukemia. Lastly, functional experiments on the short isoform, which was previously annotated as a non-coding transcript but potentially encodes a truncated protein containing only a CARD domain, would be necessary to establish a functional consequence of that isoform.

More broadly, our case example of SP140 adds to a growing list of examples of genes that become aberrantly expressed due to TE-derived promoters and enhancers, including CSF1R [43], FABP7 [42], IRF5 [44], and others [45,106–109]. Further, in primates, a paralog of SP140, SP140L, is adjacent to SP140 in the genome. While SP140L is not regulated by the intronic LTR8B, there is another intronic TE that drives cancer-specific expression in different contexts (Genbank accession BC004921.2). In conclusion, our genomic analysis suggests that the LTR8B family of ERVs is regulated by the AP1 transcription factor complex, consistent with frequent AP1 and the MAPK pathway hyperactivation seen in many cancers [110,111]. Therefore, our study supports the model where the dysregulated epigenetic environment of cancer cells leads to the aberrant activation of TE-derived regulatory elements, driving aberrant expression of genes like *SP140* that contribute to cancer pathogenesis.

Materials and Methods

Cell culture

HT1080 cells were routinely grown at 37C in 5% CO2 on coated plastic 10 cm dishes (3,000,000-5,000,000 cells) in DMEM + GlutaMAX (Gibco #10565018) supplemented with 1X penicillin-streptomycin (Gibco #15140122) and 10% fetal bovine serum (Gibco #10437010). The cell identity was validated by Anschutz Medical Campus and routinely checked for mycoplasma contamination. For dCas9-HT1080 selection the media was supplemented with 8 ug/mL Blasticidin S HCl (Gibco #A1113903), and for gRNA-HT1080 selection the media was supplemented with 1 ug/mL Puromycin (Gibco #A1113803;). Transfections were performed

using the Neon transfection system (ThermoFisher Scientific) according to the manufacturing instructions.

CRISPR-mediated silencing of LTR8B

For CRISPR-mediated silencing (CRISPRi) of the intronic SP140 LTR8B element, a HT1080 dCas9-KRAB-MeCP2 stable line was first generated using the PiggyBac system (System Bioscience). The PiggyBac donor plasmid, PB-CAGGS-dCas9-KRAB-MeCP2 was co-transfected with (0.5 ug) of Super PiggyBac transposase expression vector into HT1080 cells. The PB-CAGGS-dCas9-KRAB-MeCP2 construct was a gift from Alejandro Chavez & George Church (Addgene plasmid # 110824). 24 hours post-transfection, cells were treated with 8 ug/mL Blasticidin to select for integration of the dCas9 expression cassette, and selection was maintained for 10 days. CRISPR gRNAs specific to the DNA elements of interest (i.e., 0 predicted off-target sequences) were selected using pre-computed CRISPR target guides available on the UCSC Genome Browser hg38 assembly, and complementary oligos were synthesized by Integrated DNA Technologies. Complementary oligos were designed to generate BstXI and BlpI overhangs for cloning into PB-CRISPRia, a custom PiggyBac CRISPR gRNA expression plasmid based on the lentiviral construct pCRISPRia (a gift from Jonathan Weissman, Addgene plasmid # 84832). Complementary gRNA-containing oligos were hybridized and phosphorylated in a single reaction, then ligated into a PB-CRISPRia expression plasmid linearized with BstXI and BlpI (New England Biolabs (NEB)). NEB Stable Competent E.coli cells (#C3040H) were transformed with 2 uL of each ligation reaction and resulting colonies were selected for plasmid DNA isolation using the ZymoPure Plasmid miniprep kit (Zymo

Research). Each cloned gRNA sequence-containing PB-CRISPRia plasmid was verified by Sanger sequencing (Quintara Bio).

To generate CRISPRi stable lines, PB-CRISPRia gRNA plasmids were co-transfected with the PiggyBac transposase vector into the HT1080 dCas9-KRAB-MeCP2 polyclonal stable line. 24 hours post-transfection, cells were treated with 1 ug/mL Puromycin along with Blasticidin to select for integration of the sgRNA expression cassette(s) in dCas9 HT1080s. Selection was maintained for 5 days before transcriptional analyses.

RNA-seq

Sequencing libraries were prepared from RNA harvested from dCas9-HT1080s with or without 10 U/mL of IFNβ treatment (Proteintech #HZ-1298) for 4 hours and with control or LTR8B-SP140 KD transfection replicates. RNA from HT1080 was harvested using Zymo Quick-RNA Miniprep Plus Kit (Zymo #R1057). PolyA enrichment and library preparation was performed using the KAPA BioSystems mRNA HyperPrep Kit (KAPA #KK8581) according to the manufacturer's protocols. Briefly, 500 ng of RNA was used as input, and KAPA BioSystems single-index (KAPA #KK8700) or unique dual-index adapters (KAPA #KK8727) were added at a final concentration of 9 nM. The purified, adapter-ligated library was amplified for 11 cycles following the manufacturer's protocol. The final libraries were quantified using Qubit dsDNA High Sensitivity and TapeStation 4200 HSD5000. Libraries were pooled and sequenced on an Illumina NovaSeq 6000 (University of Colorado Genomics Core) as 150bp paired-end reads.

RNA-seq data analysis

Adapters and low-quality reads were trimmed using BBDuk v38.05 with arguments 'ktrim=r k=23 mink=11 hdist=1 qtrim=r trimq=10 tpe tbo'. Transcript quantification was conducted against GENCODE v34 using Salmon v0.13.1 with options '--libType A --validateMappings --rangeFactorizationBins 4 --gcBias'.

Differential expression analysis

Differentially expressed genes were called using DESeq2 v1.38.3 with design '~genotype+treatment+genotype:treatment'. For the CRISPRi differential expression analysis, normalized count data is derived from DESeq2 comparisons between 1) SP140-KD (4 replicates) vs GFP-gRNA control HT1080 samples (2 replicates); 2) IFNβ treated SP140-KD (4 replicates) vs IFNβ treated GFP-gRNA control HT1080 samples (2 replicates); and 3) IFNβ treated GFP-gRNA control HT1080 samples (2 replicates) vs GFP-gRNA control HT1080 samples (2 replicates). Genes with zero counts across all samples were removed.

Junction count analysis

Recount2

Using the 'snapcount' package in R (v4.2.2) (35), the following code was used to extract samples IDs (rail IDs) for SP140, 'qb.srva2 <- QueryBuilder(compilation = "srav2", regions = "SP140")'. Depending on the data source wanted, "srav2" or "gtex" can be used. To output the data frame of the SP140 junction query, 'sp140.srva2.jx.df <- query_jx(qb.srva2, return_rse = FALSE)' was

used. To filter for the exact junction of interest, 'sp140.srva2.jx.df %>% filter(start == 230230529 & end == 230237082)' was used for LTR8B-Exon2 junction, and, 'sp140_WT.srva2.jx.df = sp140.srva2.jx.df %>% filter(start == 230225904 & end == 230237082)' was used for canonical SP140 Exon1-Exon2 junction. The srva2 or gtex info table was read in from (http://snaptron.cs.jhu.edu/data/srav2/) , merged with junction rail IDs, and then reads were normalized by (coverage/spots) * 1e9.

RJunBase

Junction ID, SP140_LS008 (chr2:230230528|230237083:+), was selected for in the RJunBase website (http://www.rjunbase.org/) . 'Expression DIY' was selected followed by datasets BLCA (Bladder urothelial carcinoma), HNSC (Head and neck squamous cell carcinoma), LUAD (Lung adenocarcinoma), SARC (Sarcoma), and SKCM (Skin cutaneous melanoma).

Cancer Cell Line Encyclopedia (CCLE) analysis

Fastq ftp and md5sum info for 1,019 cancer cell lines were taken from the European Nucleotide Archive (https://www.ebi.ac.uk/) (Bioproject PRJNA523380). Salmon quant (v0.13.1) was used on fastq files with options '--libType A --validateMappings --rangeFactorizationBins 4 --gcBias -10000' and a custom hg38 gencode (v39) GTF file containing only 4 SP140 isoforms: 1) SP140-Canonical, 2) LTR8B-SP140-Long, 3) LTR8B-SP140-Short, and 4) SP140-Canonical-Short stopping at the LTR8B-SP140-Short stop codon. The SP140-Canonical-Short isoform was observed at low amounts in some bam reads from TCGA patients, but negligible expression from

salmon quant output. TPM info for these 4 isoforms was then extracted from all salmon quant output and organized into.

CUT&RUN

Libraries were prepared from dCas9-HT1080 cells without gRNA selection. Approximately 500K viable cells were used for each CUT&RUN reaction, and pulldowns were generated following the protocol from (63). All buffers were prepared according to the "High Ca2+/Low Salt" method using digitonin at a final concentration of 0.05%. The following antibodies were used at the noted dilutions: rabbit anti-mouse IgG (1:100; Abcam #ab46540), rabbit anti-H3K27ac (1:100; Abcam #ab4759). pAG-MNase was added to each sample following primary antibody incubation at a final concentration of 700 ng/mL. Chromatin digestion, release, and extraction was carried out according to the standard protocol. Sequencing libraries were generated using the KAPA BioSystems HyperPrep Kit according to the manufacturer's protocol with the following modifications: Freshly diluted KAPA BioSystems single-index adapters were added to each library at a final concentration of 9 nM. Adapter-ligated libraries underwent a double-sided 0.8X/1.0X cleanup using KAPA BioSystems Pure Beads. Purified, adapter-ligated libraries were amplified using the following PCR cycling conditions: 45 s at 98°C, (15 s at 98°C, 10 s at 60°C)*14, 60 s at 72°C. Amplified libraries underwent two 1X cleanups using Pure Beads. The final libraries were quantified using Qubit dsDNA High Sensitivity and TapeStation 4200 HSD5000. Libraries were pooled and sequenced on an Illumina NovaSeq 6000 (University of Colorado Genomics Core) as 150 bp paired-end reads.

CUT&RUN data analysis

Adapters and low-quality reads were trimmed using BBDuk v38.05 using options 'ktrim=r k=23 mink=11 hdist=1 tpe tbo qtrim=r trimq=10'. Trimmed reads were aligned to the hg38 assembly using BWA-MEM v0.7.15, and only uniquely mapping reads with a minimum MAPQ of 10 were retained. Fragments aligning with the mitochondrial genome were removed. Peak calling was performed using complete and size subsetted alignment files with MACS2 v2.1.1 with paired-end options '--format BAMPE --pvalue 0.01 --SPMR -B --call-summits'. Bigwig files were prepared from the MACS2 normalized bedgraph files using bedGraphToBigWig v4.

HiBiT tagging

To detect the presence of LTR8B-SP140-Long transcripts/proteins in HT1080 cells, we followed a CRISPR knock-in protocol to endogenously insert a HiBiT tag (Promega (64)) at the 3' terminus of SP140. Briefly, dCas9-HT1080 cells were electroporated with the IDT Alt-R CRISPR-Cas9 system (recombinant Cas9, tracrRNA, crRNA) with a single-stranded oligodeoxynucleotides (ssODN) donor template and a gRNA targeting the 3' end of SP140. The ssODN template was designed to append a terminal HiBiT tag upstream of the endogenous stop codon. Both the ssODN and gRNA were synthesized by IDT. Clonal lines were isolated using the limited dilution method in a 96-well plate format, and heterozygous and homozygous clones were identified and screened using the Nano-Glo HiBiT Lytic Detection System (Promega #N3030). We confirmed in-frame insertions by PCR, Sanger sequencing, and western blot.

Protein extraction and western blot analysis

Proteins were extracted from wild-type and SP140-HiBiT HT1080 cells using RIPA lysis buffer containing a protease inhibitor. Briefly, 1 million cells were lysed in 100 µl RIPA buffer by vortexing and then incubated on ice for 15 minutes. Lysates were then sonicated 6 times, for 2 seconds at 30-second intervals, at 50% amplitude using a Branson Digital Sonifier. Lysates were then incubated on ice for a further 15 minutes, and then centrifuged at 13,000 x g for 5 minutes at 4°C to remove cellular debris. The supernatant was retained, and the protein concentration was determined using a BCA assay (Thermo Scientific). Protein lysates were denatured in 1X LI-COR orange loading buffer and 10% β-mercaptoethanol at 90°C for 5 minutes. For western blot analysis, 50 µg of protein was loaded on a 4-12% Bis-Tris SDS-page gel that was run at 200V for 35 minutes. Protein was transferred to a PVDF membrane (0.45 µm pore) using the Thermo Fisher Miniblot modules at 20V for 1 hour. Membranes were then blocked for 1 hour at room temperature, using an Intercept (TBS) Blocking Buffer (LI-COR). Membranes were incubated overnight at 4°C, using either an anti-HiBiT (1:500; Promega) or anti-β-actin (1:20,000; CST) primary antibody diluted in blocking buffer. Primary antibodies were removed by washing the membrane 3 times in 1X TBST for 5 minutes each. Membranes were then incubated for 1 hour at room temperature with a Donkey anti-mouse 680RD secondary antibody (LI-COR), diluted in blocking buffer (anti-HiBiT = 1:10,000 and anti- β-actin = 1:20,000). The wash steps were repeated and then the membrane was imaged using the LI-COR Odyssey CLx Infrared Imaging System using the recommended settings.

Immunofluorescence protocol

Approximately 20,000 wild-type and SP140-HiBiT HT1080 cells were seeded in SensoPlates Plus 96 well plates (Greiner Bio-One) and grown in DMEM for 2-3 days. Once cells reached 80% confluency, they were serum starved for 4 hours, washed with 1X DPBS, and then fixed in 4% paraformaldehyde for 15 mins. Fixed cells were gently washed three times with 1X DPBS over 5 minutes and then cells were permeabilized in ice-cold Phosflow Perm Buffer III (BD Biosciences) for 10 minutes at -20°C. After repeating the wash steps, the permeabilized cells were blocked for 1 hour with 5% goat-serum-based buffer (1X PBS, 0.3% Triton X-100). Cells were then incubated for 2 hours with the anti-HiBiT antibody (Promega) diluted in blocking buffer (1:100). After repeating the wash steps, the cells were incubated for 1 hour with a Goat Anti-mouse secondary antibody conjugated to an Alexa Fluor® 647 (Abcam) diluted in blocking buffer (1:1000). After washing off the secondary antibody, the cell nuclei were stained with DAPI (1 μg/ml) for 10 minutes followed by a final three washes. Cells were imaged using a Nikon Spinning Disc Confocal Yokogawa CSU X1 microscope using two different lasers [laser intensity: 405 nm (15%), 640 nm (20%)], EM Gain 10MHz, 300 ms exposure, controlled by the NIS Element v5.42.03 software.

Cell viability assay

dCas9-HT1080 cells were transfected with SP140 KD or GFP control, selected with Blasticidin and Puromycin for 5 days, and then harvested. 5,000 cells were seeded into 96-well dishes (Greiner Bio-One #655086). The next day, cells were treated with 0, 10, 100, 1000 U/mL of IFNβ in triplicates. For each day, a 96-well plate was assayed through a luminescence assay of

ATP (CellTiter-Glo 2.0 Assay, Promega #G9242) following the manufacturer's instructions. A two-sample equal variance t-test was used to assess significance.

Transcriptome Assembly (StringTie)

RNA-Seq fastq files from publicly available HT1080 (GEO: GSE68109) and U2OS (GEO: GSE66789) were used as input for StringTie (65) to assemble transcripts without a reference annotation guide. Adapters and low-quality reads were trimmed using BBDuk v38.05 with arguments 'ktrim=r k=23 mink=11 hdist=1 qtrim=r trimq=10 tpe tbo' and mapped to hg38 using hisat2 v2.1.0 with options '--rna-strandness RF --no-softclip --dta'. Aligned fragments with an alignment score of less than 10 were removed with samtools v1.10. StringTie (v1.3.3b) was then used with options '--rf'.

Assessment of LTR8B-SP140 isoform expression levels by RT-qPCR

HT1080 cells were lysed in 300µl of RNA lysis buffer (Zymo Research #R1060-1-50), and were stored at -80°C or immediately used for RNA extraction. RNA extraction was performed using the Quick-RNA MiniPrep kit (Zymo Research #R1054) following the manufacturer's instructions. A NanoDrop One spectrophotometer (Thermo Fisher Scientific) was used to determine RNA concentration and quality. A list of primers used for RT-qPCR and cycling conditions are provided in. RNA expression levels were quantified using the Luna Universal One-Step RT-qPCR Kit (New England Biolabs #E3005L) according to the manufacturer's instructions. In brief, for each reaction, 25ng of RNA was combined with 5µl 2× Luna Universal One-Step Reaction Mix, 0.5µl 20× Luna WarmStart RT Enzyme Mix, 0.4µl 10µM forward

primer, and 0.4µl 10µM reverse primer. Reactions were amplified using a CFX384 Touch Real-Time PCR Detection System (Bio-Rad). On-target amplification was assessed by melt curve analysis. Each sample was run either in technical duplicate or triplicate. RT-qPCR result values were analyzed using the ΔCq expression method (2^(-ΔCq)) normalizing Ct values of target genes to the Ct value of the CTCF housekeeping gene per each sample/replicate. A paired t-test was used to assess significance.

Cistrome analysis

From the CUT&RUN data analysis MACS2 output, the HT1080 H3K27ac narrowPeak files were used as input for the bedtools intersect (v2.28.0) (66) against an LTR8B bed file containing coordinates of the 1,730 LTR8B loci. This generated a bed file containing the 70 LTR8B loci that overlapped with H3K27ac. A background file of the 1,660 LTR8B loci not overlapping with H3K27ac was also generated using the same method. These bed files were used as input in the cistrome toolkit data browser (http://dbtoolkit.cistrome.org/) looking for transcription factors and chromatin regulators in human hg38.

MEME analysis

Motif analysis of LTR8B loci that overlap with H3K27ac was performed using the MEME suite (v5.1.0) (47) in differential enrichment mode. The bed files from the cistrome analysis were converted into fasta files using 'bedtools getfasta'. H3K27ac-overlapped LTR8B sequences (n=70) were used as input against a background set of non-overlapped LTR8B sequences (n=1,660), using default settings except for the number of motifs to find (5). Each discovered

motif was searched for similarity to known motifs using the JASPAR 2018 non-redundant DNA database with TomTom (v5.5.5).

Appendix A – Finding Oncogene-Induced Senescent Specific Transposable Elements and picking out SP140

Chronic inflammation is a prominent feature of aging/age-related diseases and is driven in part by senescent cells that secrete pro-inflammatory molecules [48]. How and why senescent cells transition into an inflammatory state is poorly understood at the molecular level, with work done by De Cecco et al [51] proposing TE transcripts mimicking viral RNA in the cytoplasm activating an IFN response within the cell. However, TEs often act as non-coding enhancer elements, particularly in the context of inflammatory pathways [41], which is an avenue of research that has not been explored for senescent cells. I hypothesize that reactivation of TEs may underlie gene expression patterns specific to an inflammatory state.

Senescent cells, cells with halted cell division, have been found to have epigenetically permissive chromatin - opening a 'Pandora's box' full of TEs newly accessible to cellular machinery. These newly accessible TEs may act as senescent-specific enhancer elements that contribute to chronic inflammation. I aim to use genomic and experimental methods to investigate the potentially major role for TEs as regulatory elements that induce inflammation associated genes in senescent cells. This would unveil gene dysregulation as a novel mechanism. If enhancer TEs are causative to increased expression of inflammatory associated genes, enhancer TEs can be a viable therapeutic target for reducing age related morbidity and further cement the importance of TEs in cell function and pathology.

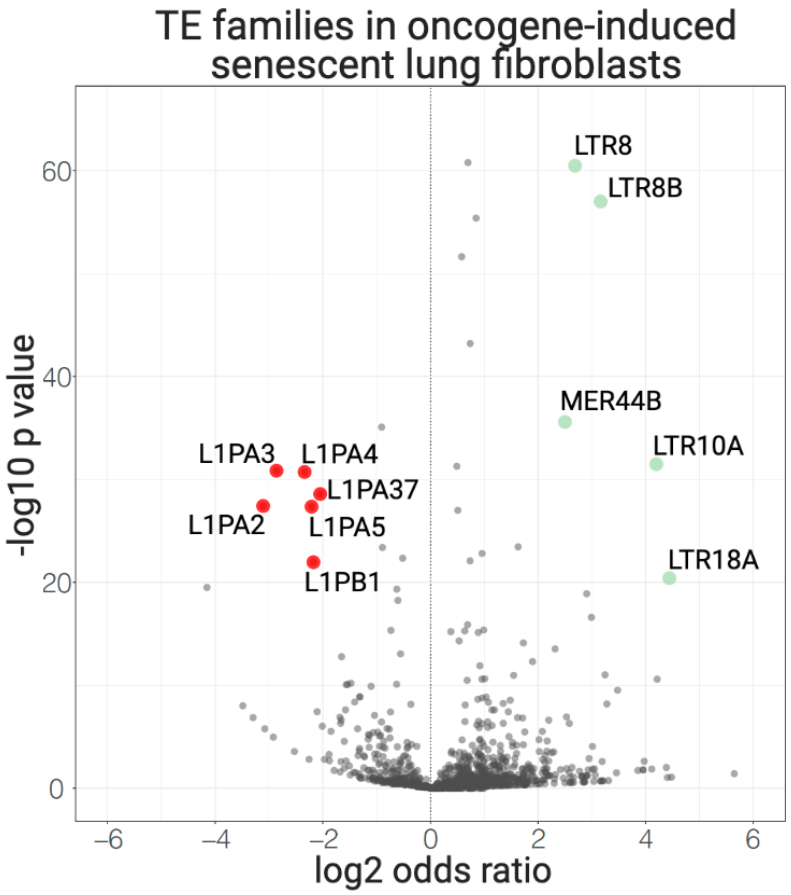
I re-analyzed chromatin immunoprecipitation (ChIP) sequence and RNA-sequence data from Tasdemir et al [112] looking at human lung fibroblast cells (IMR90). Tasdemir et al took data from normal proliferating IMR90 cells, quiescent IMR90 (a non proliferative cell type which is

not inflammatory; induced by culturing IMR90 cells at .1% FBS for 4 days), and oncogene-induced senescent (OIS) IMR90 cells (made by retroviral-mediated expression of H-Ras^{V12} [55]).

I first wanted to know if there were OIS IMR90 specific TEs. To find this out, I extracted the H3K27ac ChIP-seq data for the 3 IMR90 cell states. H3K27ac ChIP-seq data is important for this step because of H3K27ac is a mark in the genome for active enhancers and transcription start sites, allowing me to detect TEs that overlap with H3K27ac as active enhancers or transcription start sites. I then extracted out all the OIS IMR90 specific TEs which overlap with H3K27ac using an R-studio package, DeSeq2, as a differential peak analysis. I then took the OIS IMR90 specific TEs that overlap H3K27ac and ran them through a colocalization analysis, GIGGLE [94], to identify TE families that are overrepresented in OIS IMR90 cells that overlap with H3K27ac as targets (Figure A1.1). This list narrows downs the hundreds of thousands of TEs in the human genome as possible gene regulators for OIS IMR90 to just a few thousand.

Next, I picked out the top two TE families from (Figure A1.1), LTR8 and LTR8B, to narrow the TEs even more. These two families are closely related, with LTR8B being a subfamily of LTR8, meaning it represents the divergence and forking of LTR8 into its own family [113,114].

I then visually verified the TEs that overlapped with H3K27ac in LTR8B by a heatmap (Figure A1.2) and found similar results for LTR8. Figure A1.2 verifies there are more H3K27ac peaks in OIS IMR90 cells, and narrows my focus even further to the OIS specific H3K27ac LTR8B loci.



A1.1. Oncogene-Induced Senescent IMR90 specific overrepresented TE families as possible enhancers or transcription start sites. Volcano plot of differentially represented peaks using DeSeq2 between OIS IMR90 and quiescent IMR90 cells. Each dot represents a TE family. Overrepresented TEs in OIS IMR90 are in green and underrepresented TEs are in red. 2 replicates are used.

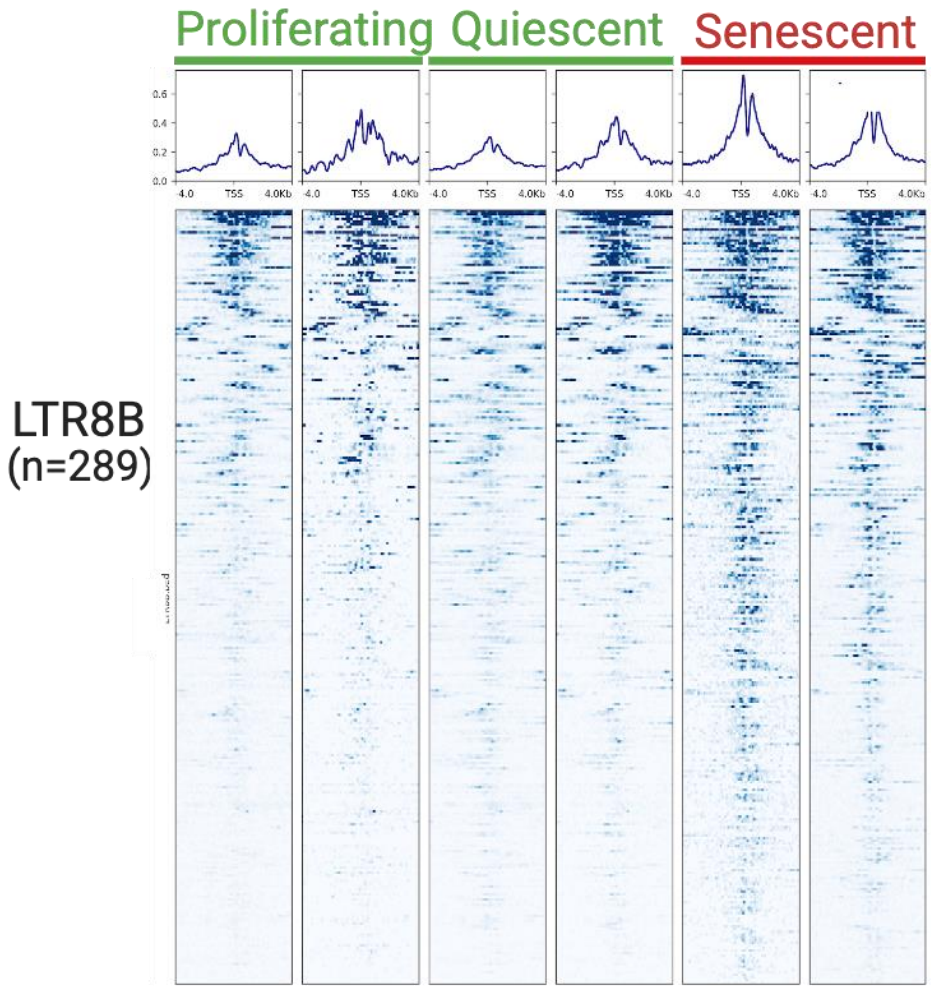


Figure A1.2. LTR8B loci that line up with H3K27ac peaks. Heat map of H3K27ac peaks at LTR8B loci. Each row specifies an LTR8B loci within the human genome, and the blue signal for each row represents H3K27ac binding at that region. There are 289 LTR8B loci with H3K27ac overlap between the samples out of 1730 total LTR8B loci in the human genome. 2 replicates are used for each IMR90 cell condition.

Next, I took the OIS specific H3K27ac LTR8B loci and asked if any of those loci overlapped or were within enhancer acting distance (50kb) of genes related to inflammation, SASP, immune dysregulation, apoptosis, histone modifications, and transcription factors (Figure A1.3). This narrowed my list of LTR8 and LTR8B loci even further for studying its effect on gene expression and inducing an inflammatory milieu. Some notable genes in this list were C3 [115], IL33 [116], TIMP1 [117], and AHR [118] – all genes important in mediating immune response and inflammation. However, the gene I moved forward for this thesis was SP140 for two reasons: 1) SP140 was the only gene in this list that directly overlapped LTR8B, making it more likely that the LTR8B would play a role in regulation; and 2) SP140 is linked to regulating inflammation [52–54].

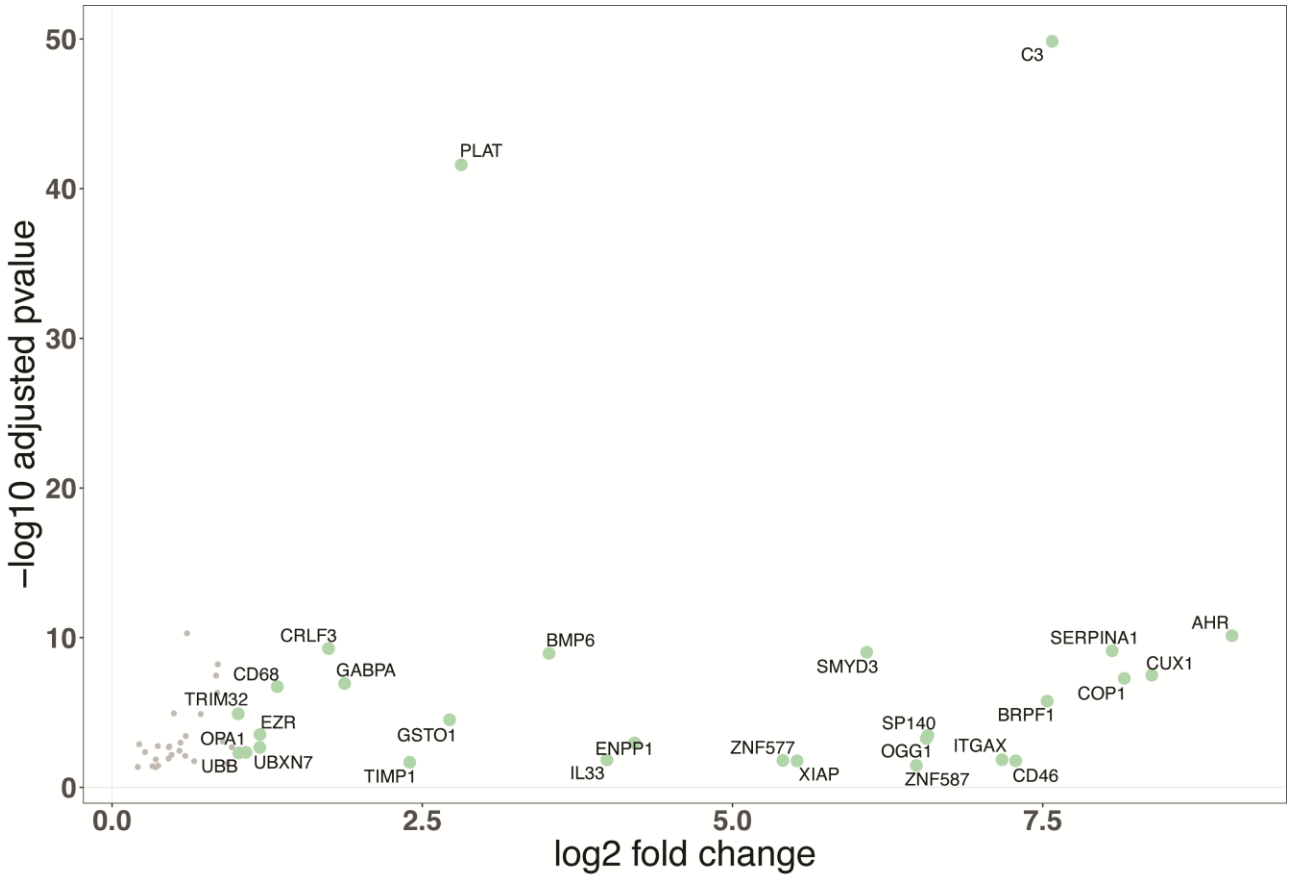


Figure A1.3. Inflammatory/disease related genes within 50kb of OIS IMR90 specific LTR8B loci.
Volcano plot of differentially regulated genes (upregulated) between OIS IMR90 and quiescent and proliferating IMR90 cells within 50kb of OIS IMR90 specific LTR8 and LTR8B. Each dot represents an upregulated gene.

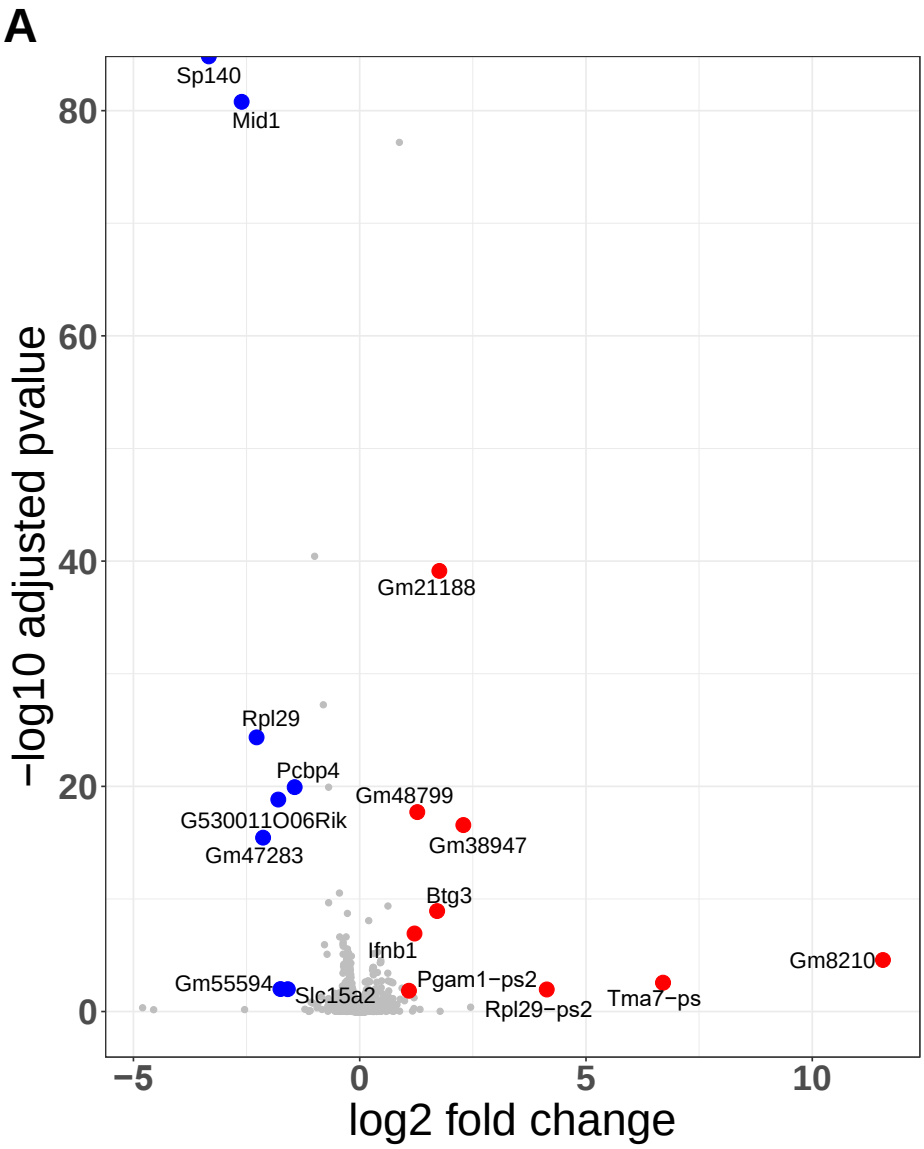
Appendix B - Canonical Sp140 mechanism in mice

A previous study Mehta et al. 2017 [64], showed that in human macrophages activated by IFN-gamma alone or IFN-gamma plus LPS, the SP140 protein binds to DNA mostly at gene promoters and lineage-inappropriate genes marked by H3K27me3. This interaction plays a crucial role in controlling the gene expression patterns of macrophages, both under normal conditions and when they respond to cytokines or microbial infections. SP140 was most abundantly associated with HOX genes, especially HOXA9, which is known to maintain a stem-like state in hematopoietic stem cells and to block macrophage development. Thus, they suggest that SP140 helps define macrophage identity and function by repressing genes that are inappropriate for the macrophage lineage, especially during responses to cytokines and pathogens.

In collaborating with Dr. Witt and Dr. Vance at UC - Berkeley, they have conducted SP140 KO in mice along with IFN stimulation (DMXAA) to look further into the mechanism of canonical SP140. I analyzed RNA-seq, CUT&RUN, and ATAC-seq data for these conditions. What they found and what I confirmed with analysis is that SP140 negatively regulates mRNA stability of the type I interferon (IFN-I) gene Ifnb1.

SP140 is a transcriptionally repressive epigenetic reader that lacks RNA-binding domains. Kristin hypothesized that SP140 indirectly regulates Ifnb1 mRNA stability by repressing the transcription of an unknown factor. To identify this factor, Dr. Witt generated RNA-seq data from DMXAA-treated B6 and Sp140–/– bone marrow-derived macrophages (BMMs), as well as DMXAA-treated Ifnar$^{-/-}$ and Sp140$^{-/-}$Ifnar$^{-/-}$ BMMs to account for the confounding effects of elevated IFN-I signaling through IFNAR in Sp140$^{-/-}$ BMMs. Few genes were differentially expressed between DMXAA-treated Ifnar$^{-/-}$ and Sp140$^{-/-}$Ifnar$^{-/-}$ BMMs

besides Ifnb1 and Sp140 (Figure A2.1-A). Interestingly, only two DEGs correlated with Ifnb1 upregulation across RNA-seq datasets: 1) Sp140, which is downregulated in Sp140 knockout cells, and 2) the poorly annotated gene, Gm21188, which was upregulated in SP140 knockout cells (Figure A2.1-B). Gm36079, a copy of Gm21188 encoding an identical protein, was only significantly differentially expressed in Sp140–/– BMMs treated with DMXAA, suggesting IFNAR signaling in the absence of SP140 is required for robust induction (Figure A2.1-B). As upregulation of Gm21188/Gm36079 and Ifnb1 strikingly correlated in Sp140–/– cells, Dr. Witt hypothesized that Gm21188/Gm36079 encode a novel positive regulator of Ifnb1 mRNA stability, which is transcriptionally repressed by SP140.


A
Sp140
Mid1
Gm21188
Rpl29
Pcbp4
Gm48799
G530011O06Rik
Gm47283
Gm38947
Btg3
Ifnb1
Gm8210
Gm55594
Slc15a2
Pgam1−ps2
Rpl29−ps2
Tma7−ps
80
60
40
20
0
−log10 adjusted pvalue
−5
0
5
10
log2 fold change

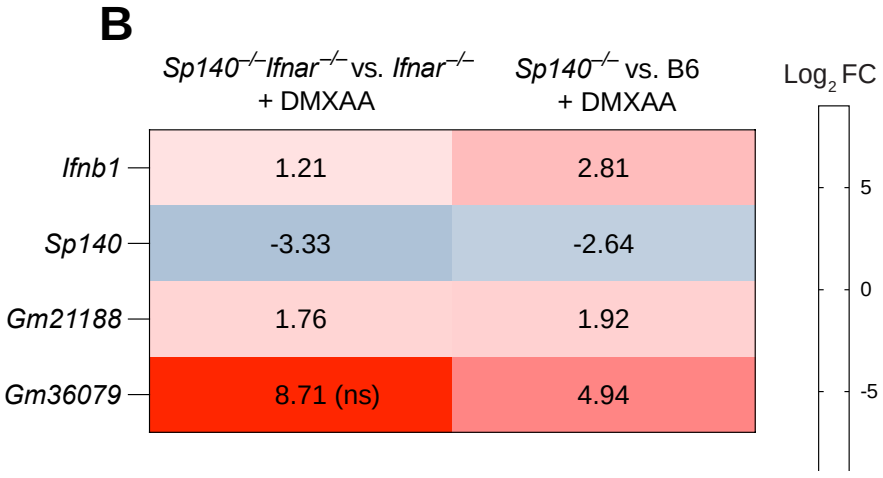

B
Sp140−/−Ifnar−/− vs. Ifnar−/−
+ DMXAA
Sp140−/− vs. B6
+ DMXAA
Log2 FC
Ifnb1
Sp140
Gm21188
Gm36079
1.21
2.81
-3.33
-2.64
1.76
1.92
8.71 (ns)
4.94
5
0
-5

Figure A2.1. Gm21188/Gm36079 are repressed by SP140 and correlate with increased Ifnb1 transcript in Sp140–/– cells. **A)** Volcano plot of differentially expressed genes (DEGs) from RNA-seq of DMXAA-treated Sp140–/–Ifnar–/– vs. Ifnar–/– BMMs. Red genes are upregulated in Sp140–/–Ifnar–/– BMMs with log2 fold change > 1 and adjusted p value < 0.05. Blue genes are downregulated in Sp140–/–Ifnar–/– BMMs with log2 fold change > -1 and adjusted p value < 0.05. **B)** Table of 1) the 3 DEGs (Ifnb1, Gm21188, Sp140) shared across RNA-seq datasets of DMXAA-treated Sp140–/–Ifnar–/– vs. Ifnar–/– and Sp140–/– vs. B6 BMMs, and 2) Gm36079, which was significantly upregulated in DMXAA-treated Sp140–/– vs. B6, but not Sp140–/–Ifnar–/–vs. Ifnar–/– BMMs (ns = not significant). Cells are colored by log2 fold change (Figure A2.1-B made by Dr. Witt).

SP140 robustly bound and repressed chromatin accessibility at the Gm21188/Gm36079 locus (Figure A2.2). From the list of upregulated genes in DMXAA-treated Sp140–/–Ifnar–/– vs. Ifnar–/– BMMs, only Gm21188 was both bound by SP140 (CUT&RUN) and showed increased chromatin accessibility upon SP140 knockout (ATAC-seq) (Figure A2.2). SP140 binds a ~10 kb region encompassing Gm36079/Gm21188, and negatively regulates chromatin accessibility at Gm36079/Gm21188 gene loci (Figure A2.2). These results further support Dr. Witt's hypothesis that Gm21188/Gm36079 are repressed by SP140 and encode a novel positive regulator of Ifnb1 transcript stability.

Further, consistent with previous SP140 ChIP-seq results from Mehta et al [64], we confirmed that SP140 generally binds and represses chromatin opening at genes involved in development (Figure A2.3-A), like Hoxa9 (Figure A2.3-B), however, unlike Gm21188/Gm36079, SP140 binding at the Hoxa9 loci does not affect transcription. SP140-binding also correlated with the transcriptionally repressive histone mark H3K27me3 in publicly available ChIP-seq datasets (Figure A2.3-C). Consistent with Dr. Witt's hypothesis that SP140 indirectly represses Ifnb1 mRNA stability, SP140 does not bind and regulate chromatin accessibility at the Ifnb1 gene (Figure A2.4-A) or known regulatory elements [119–122] (Figure A2.4-B).

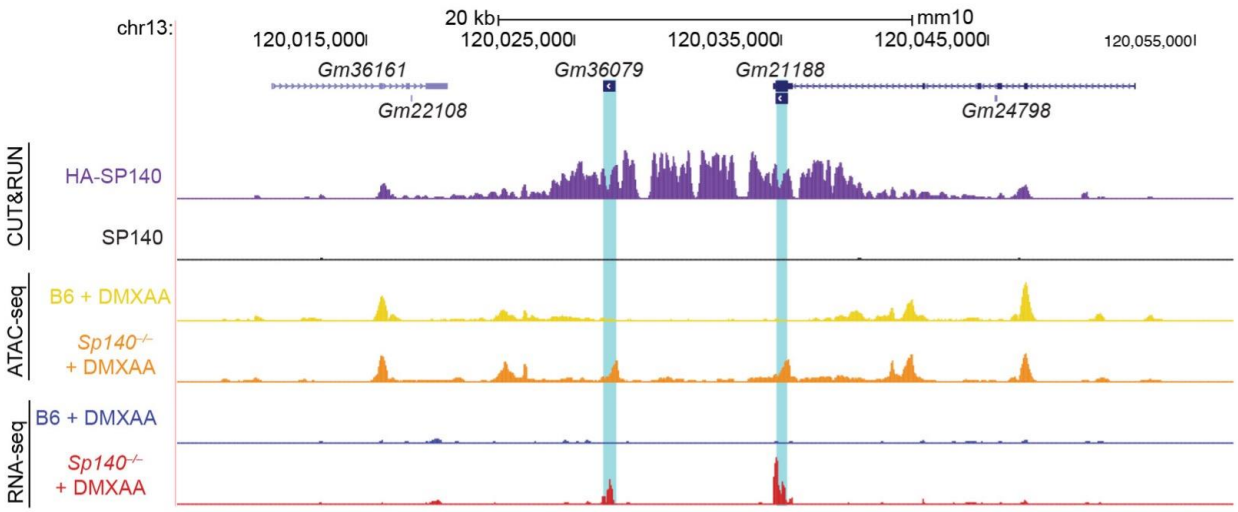


Figure A2.2. SP140 binds at Gm36079 and Gm21188. Alignment of reads at Gm21188/Gm36079 locus from anti-HA CUT&RUN for DMXAA-treated BMMs transduced with HA-SP140 or SP140, and ATAC-seq/RNA-seq of DMXAA-treated B6 and Sp140–/– BMMs. Alignments were visualized in the UCSC genome browser.

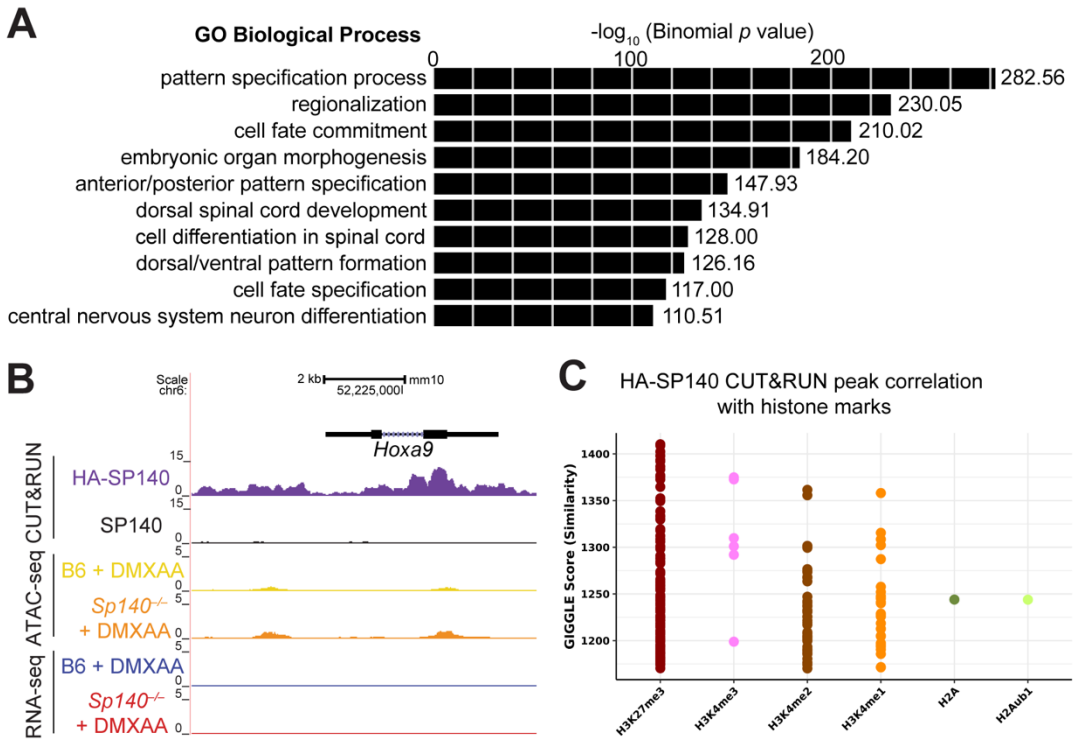


Figure A2.3. SP140 predominantly binds and represses chromatin accessibility at genes involved in development. **A)** Top 10 GO terms for genes bound by HA-SP140 in anti-HA CUT&RUN. **B)** Volcano plot of differentially accessible ATAC-seq peaks in Sp140–/– BMMs treated with DMXAA vs. B6 BMMs treated with DMXAA, filtered by genes that are also bound by HA-SP140 in anti-HA CUT&RUN. **C)** Alignment of reads from HA-SP140 or untagged SP140 anti-HA CUT&RUN, and RNA seq/ATAC-seq of Sp140–/– and B6 BMMs treated with DMXAA at Hoxa9.

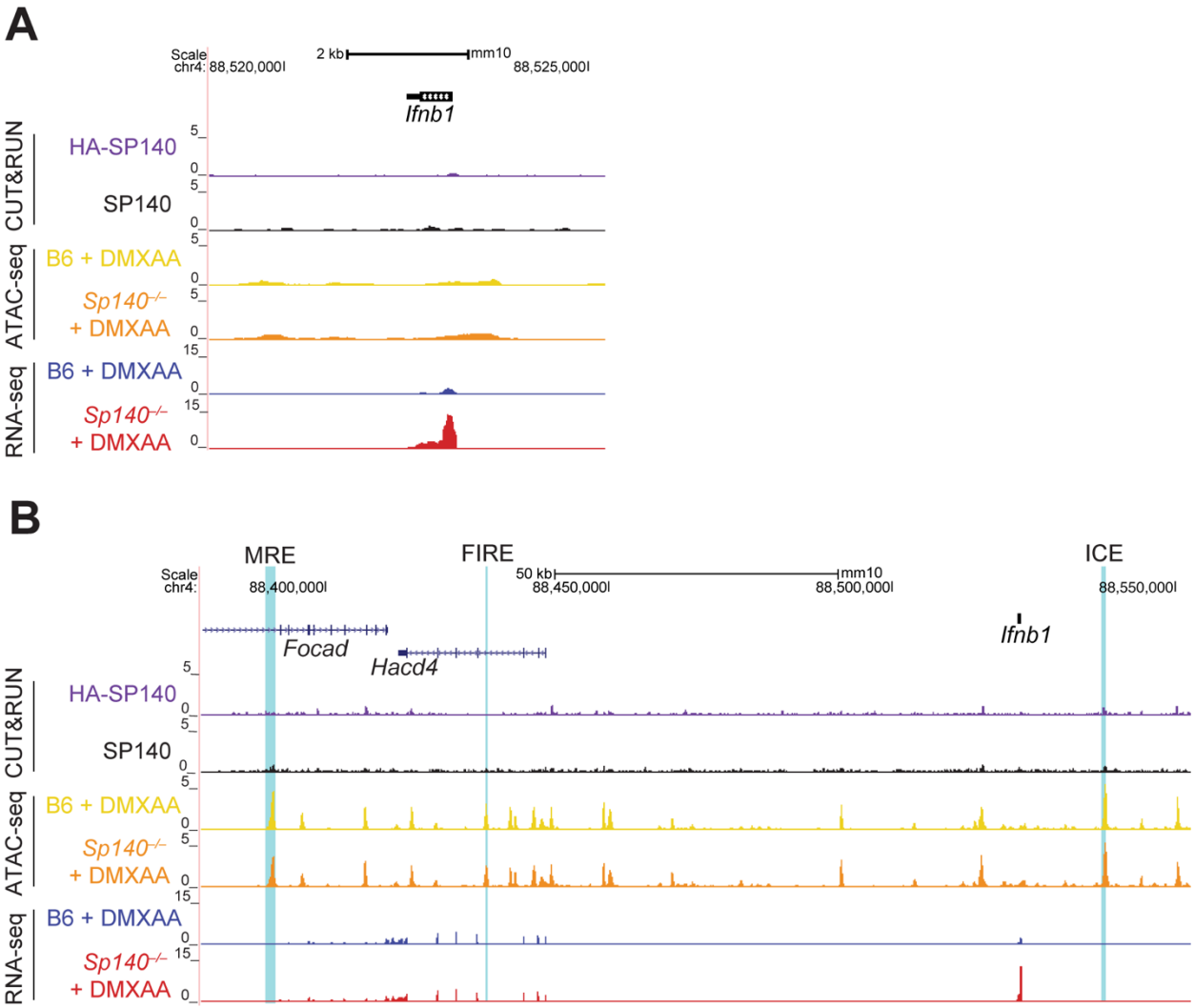


Figure A2.4. SP140 does not bind the Ifnb1 locus or known regulatory elements. **A)** Alignment of reads from HA-SP140 or untagged SP140 anti-HA CUT&RUN, ATAC-seq of Sp140–/– and B6 BMMs treated with DMXAA, and RNA-seq of Sp140–/– and B6 BMMs treated with DMXAA at Ifnb1. **B)** Alignment of reads from HA-SP140 or untagged SP140 anti-HA CUT&RUN, and ATAC-seq/RNA-seq of Sp140–/– and B6 BMMs treated with DMXAA, at the Ifnb1 regulatory elements ICE36,37, FIRE35, and the MRE34.

Conclusion & Future Directions

From our studies, we were able to identify two ERV derived isoforms of SP140 in a novel context – non-immune cancer cells. The long isoform (*LTR8B-SP140-Long*) is completely novel and unidentified, while the short isoform (*LTR8B-SP140-Short*) is annotated in GENCODE [123] as a non-protein coding transcript. We were able to determine these ERV derived isoforms play a role in altering transcription (Figure 2.9) and immunomodulatory function (Figure 2.8B) protecting cells from death or proliferative arrest upon IFNB treatment possibly due to the dysregulation of NF-κB. However, further evidence is needed to determine the effects of these isoforms *in vitro* and *in vivo*.

First, there are a few experiments which would strengthen the case of LTR8B-SP140 in cancer cells functioning similarly to its canonical counterpart in immune cells. Previous studies on SP140 in human macrophages shows it is binding chromatin at HOX genes by ChIP-seq [64], preventing transcription of genes at the loci SP140 binds to. To further understand the mechanism of the *LTR8B-SP140-Long* isoform, it would be beneficial to prove that it also binds to chromatin and to see if it is affecting transcription at the loci it binds. To see if the *LTR8B-SP140-Long* isoform also binds chromatin in HT1080s, I propose implementing the HiBiT system I mentioned earlier in chapter II. The 11 amino acid HiBiT tag at the 5' end of *LTR8B-SP140-Long* protein would provide a target to pull down using the HiBiT antibody for CUT&RUN [124]. If binding occurs, I would then be able to view the binding on UCSC genome browser [125] and observe if it matches up to any of the genes from (Figure 2.9).

Further, it would be interesting to replicate the role in altering transcription (Figure 2.9) and immunomodulatory function (Figure 2.8B) in other non-immune cancer cell types such as 769P (kidney), YD15 (salivary gland), and LOXIMVI (skin). These cell lines express LTR8B-

SP140 up to ~13x greater than HT1080 and in different cell types. Due to the fold difference in expression, it would be interesting to observe if these cells would have greater interferon resistance and, therefore, a greater knockdown effect of LTR8B-SP140 through a cytotoxicity experiment with IFNB. Further, it would be interesting to observe if the same genes as (Figure 2.9) are affected across different cell types.

Another important experiment would be to parse out the function of the *LTR8B-SP140-Long* and *LTR8B-SP140-Short* isoforms. In chapter II, I have shown the long isoform forms protein aggregates in the nucleus, however, the GENCODE predicted non-coding short isoform has yet to be characterized. The short isoform is important since it contains the caspase-associated recruitment domain (CARD). CARD containing proteins have been established as key regulators of cell death and cytokine production participating in NF-κB signaling pathways associated with innate or adaptive immune response [126]. Further, CARD containing proteins are being explored as therapeutic drug targets in the treatment of cancer [127]. Given that CARD-CARD interactions mediate the formation of large signaling complexes [128], the LTR8B-SP140-Short CARD protein (if translated) could oligomerize with other CARD containing proteins affecting apoptosis and immune response. This builds up the case for making over-expression vectors of the *LTR8B-SP140-Long*, *LTR8B-SP140-Short*, and the canonical *SP140* isoform to reintroduce into HT1080 knockout cell lines. Then retest cytotoxic effects upon IFNB treatment to see if any of the single isoforms on its own has greater affect at rescuing than the other.

In conclusion, these *in vitro* experiments would further build our understanding of the robustness of LTR8B-SP140 in non-immune cancer cell lines along with understanding the mechanism behind the effects of the isoform(s).

References

1. Genomics Institute/Human Genome Center, B. Initial sequencing and analysis of the human genome. *Nature* (2001).
2. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
3. science.abj6987.pdf.
4. Hoyt, S. J. *et al.* From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. *Science* **376**, eabk3112 (2022).
5. de Koning, A. P. J., Gu, W., Castoe, T. A., Batzer, M. A. & Pollock, D. D. Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. *PLoS Genet.* **7**, e1002384 (2011).
6. Burns, K. H. & Boeke, J. D. Human transposon tectonics. *Cell* **149**, 740–752 (2012).
7. Castanera, R. *et al.* Transposable Elements versus the Fungal Genome: Impact on Whole-Genome Architecture and Transcriptional Profiles. *PLoS Genet.* **12**, e1006108 (2016).
8. Wu, C. & Lu, J. Diversification of Transposable Elements in Arthropods and Its Impact on Genome Evolution. *Genes* **10**, (2019).
9. Mouse Genome Sequencing Consortium *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
10. Gentles, A. J. *et al.* Evolutionary dynamics of transposable elements in the short-tailed opossum Monodelphis domestica. *Genome Res.* **17**, 992–1004 (2007).
11. Springer, N. M. *et al.* Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genet.* **5**, e1000734 (2009).

12. Sultana, T., Zamborlini, A., Cristofari, G. & Lesage, P. Integration site selection by retroviruses and transposable elements in eukaryotes. *Nat. Rev. Genet.* **18**, 292–308 (2017).
13. Kaestner, E. The origins of genome architecture. (2016) doi:10.5860/choice.45-0862.
14. Kazazian, H. H., Jr. Mobile elements: drivers of genome evolution. *Science* **303**, 1626–1632 (2004).
15. Cordaux, R. & Batzer, M. A. The impact of retrotransposons on human genome evolution. *Nat. Rev. Genet.* **10**, 691–703 (2009).
16. Goodier, J. L. & Kazazian, H. H., Jr. Retrotransposons revisited: the restraint and rehabilitation of parasites. *Cell* **135**, 23–35 (2008).
17. Wicker, T. *et al.* A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8**, 973–982 (2007).
18. Jurka, J., Bao, W. & Kojima, K. K. Families of transposable elements, population structure and the origin of species. *Biol. Direct* **6**, 44 (2011).
19. Feschotte, C. Transposable elements and the evolution of regulatory networks. *Nat. Rev. Genet.* **9**, 397–405 (2008).
20. Slotkin, R. K. & Martienssen, R. Transposable elements and the epigenetic regulation of the genome. *Nat. Rev. Genet.* **8**, 272–285 (2007).
21. Mills, R. E., Bennett, E. A., Iskow, R. C. & Devine, S. E. Which transposable elements are active in the human genome? *Trends Genet.* **23**, 183–191 (2007).
22. Molaro, A. & Malik, H. S. Hide and seek: how chromatin-based pathways silence retroelements in the mammalian germline. *Curr. Opin. Genet. Dev.* **37**, 51–58 (2016).
23. Liu, N. *et al.* Selective silencing of euchromatic L1s revealed by genome-wide screens for L1 regulators. *Nature* **553**, 228–232 (2018).

24. O'Donnell, K. A. & Boeke, J. D. Mighty Piwis defend the germline against genome intruders. *Cell* **129**, 37–44 (2007).

25. Yang, F. & Wang, P. J. Multiple LINEs of retrotransposon silencing mechanisms in the mammalian germline. *Semin. Cell Dev. Biol.* **59**, 118–125 (2016).

26. Smalheiser, N. R. & Torvik, V. I. Mammalian microRNAs derived from genomic repeats. *Trends Genet.* **21**, 322–326 (2005).

27. Kanellopoulou, C. *et al.* Dicer-deficient mouse embryonic stem cells are defective in differentiation and centromeric silencing. *Genes Dev.* **19**, 489–501 (2005).

28. Schmitges, F. W. *et al.* Multiparameter functional diversity of human C2H2 zinc finger proteins. *Genome Res.* **26**, 1742–1752 (2016).

29. Imbeault, M., Helleboid, P.-Y. & Trono, D. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* **543**, 550–554 (2017).

30. Yang, P., Wang, Y. & Macfarlan, T. S. The Role of KRAB-ZFPs in Transposable Element Repression and Mammalian Evolution. *Trends Genet.* **33**, 871–881 (2017).

31. Zhang, Y. *et al.* Transposon molecular domestication and the evolution of the RAG recombinase. *Nature* **569**, 79–84 (2019).

32. Cornelis, G. *et al.* Retroviral envelope gene captures and syncytin exaptation for placentation in marsupials. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E487-96 (2015).

33. Pastuzyn, E. D. *et al.* The Neuronal Gene Arc Encodes a Repurposed Retrotransposon Gag Protein that Mediates Intercellular RNA Transfer. *Cell* **173**, 275 (2018).

34. Wang, T. *et al.* Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 18613–18618 (2007).

35. Kunarso, G. *et al.* Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat. Genet.* **42**, 631–634 (2010).
36. Schmidt, D. *et al.* Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* **148**, 832 (2012).
37. Chuong, E. B., Rumi, M. A. K., Soares, M. J. & Baker, J. C. Endogenous retroviruses function as species-specific enhancer elements in the placenta. *Nat. Genet.* **45**, 325–329 (2013).
38. Jacques, P.-É., Jeyakani, J. & Bourque, G. The majority of primate-specific regulatory sequences are derived from transposable elements. *PLoS Genet.* **9**, e1003504 (2013).
39. Sundaram, V. *et al.* Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Res.* **24**, 1963–1976 (2014).
40. Fuentes, D. R., Swigut, T. & Wysocka, J. Systematic perturbation of retroviral LTRs reveals widespread long-range effects on human gene regulation. *Elife* **7**, (2018).
41. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* **351**, 1083–1087 (2016).
42. Lock, F. E. *et al.* Distinct isoform of FABP7 revealed by screening for retroelement-activated genes in diffuse large B-cell lymphoma. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E4630 (2015).
43. Lamprecht, B. *et al.* Derepression of an endogenous long terminal repeat activates the CSF1R proto-oncogene in human lymphoma. *Nat. Med.* **16**, 571–9, 1p following 579 (2010).
44. Babaian, A. *et al.* Onco-exaptation of an endogenous retroviral LTR drives IRF5 expression in Hodgkin lymphoma. *Oncogene* **35**, 2542–2546 (2016).

45. Ivancevic, A. *et al.* Endogenous retroviruses mediate transcriptional rewiring in response to oncogenic signaling in colorectal cancer. *bioRxiv* 2021.10.28.466196 (2023) doi:10.1101/2021.10.28.466196.

46. Hayflick, L. & Moorhead, P. S. The serial cultivation of human diploid cell strains. *Exp. Cell Res.* **25**, 585–621 (1961).

47. Coppé, J.-P., Desprez, P.-Y., Krtolica, A. & Campisi, J. The senescence-associated secretory phenotype: the dark side of tumor suppression. *Annu. Rev. Pathol.* **5**, 99–118 (2010).

48. López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. The hallmarks of aging. *Cell* **153**, 1194–1217 (2013).

49. Wiley, C. D. & Campisi, J. From ancient pathways to aging cells—connecting metabolism and cellular senescence. *Cell Metab.* **23**, 1013–1021 (2016).

50. Franceschi, C. & Campisi, J. Chronic inflammation (inflammaging) and its potential contribution to age-associated diseases. *J. Gerontol. A Biol. Sci. Med. Sci.* **69 Suppl 1**, S4-9 (2014).

51. De Cecco, M. *et al.* Author Correction: L1 drives IFN in senescent cells and promotes age-associated inflammation. *Nature* **572**, E5 (2019).

52. Ghiboub, M. *et al.* Modulation of macrophage inflammatory function through selective inhibition of the epigenetic reader protein SP140. *BMC Biol.* **20**, 182 (2022).

53. Ji, D. X. *et al.* Role of the transcriptional regulator SP140 in resistance to bacterial infections via repression of type I interferons. *Elife* **10**, (2021).

54. Karaky, M. *et al.* SP140 regulates the expression of immune-related genes associated with multiple sclerosis and other autoimmune diseases by NF-κB inhibition. *Hum. Mol. Genet.* **27**, 4012–4023 (2018).

55. Narita, M. *et al.* Rb-mediated heterochromatin formation and silencing of E2F target genes during cellular senescence. *Cell* **113**, 703–716 (2003).

56. Dziulko, A. K., Allen, H. & Chuong, E. B. An endogenous retrovirus regulates tumor-specific expression of the immune transcriptional regulator SP140. *Hum. Mol. Genet.* (2024) doi:10.1093/hmg/ddae084.

57. Jenuwein, T. & Allis, C. D. Translating the histone code. *Science* **293**, 1074–1080 (2001).

58. Ruthenburg, A. J., Allis, C. D. & Wysocka, J. Methylation of lysine 4 on histone H3: intricacy of writing and reading a single epigenetic mark. *Mol. Cell* **25**, 15–30 (2007).

59. Fraschilla, I. & Jeffrey, K. L. The Speckled Protein (SP) Family: Immunity's Chromatin Readers. *Trends Immunol.* **41**, 572–585 (2020).

60. Bloch, D. B., de la Monte, S. M., Guigaouri, P., Filippov, A. & Bloch, K. D. Identification and characterization of a leukocyte-specific component of the nuclear body. *J. Biol. Chem.* **271**, 29198–29204 (1996).

61. Dent, A. L. *et al.* LYSP100-associated nuclear domains (LANDs): description of a new class of subnuclear structures and their relationship to PML nuclear bodies. *Blood* **88**, 1423–1426 (1996).

62. Madani, N. *et al.* Implication of the lymphocyte-specific nuclear body protein Sp140 in an innate response to human immunodeficiency virus type 1. *J. Virol.* **76**, 11133–11138 (2002).

63. Amatullah, H. *et al.* Epigenetic reader SP140 loss of function drives Crohn's disease due to uncontrolled macrophage topoisomerases. *Cell* **185**, 3232-3247.e18 (2022).

64. Mehta, S. *et al.* Maintenance of macrophage transcriptional programs and intestinal homeostasis by epigenetic reader SP140. *Sci Immunol* **2**, (2017).

65. Oeckinghaus, A. & Ghosh, S. The NF-κB Family of Transcription Factors and Its Regulation. *Cold Spring Harb. Perspect. Biol.* **1**, a000034 (2009).

66. Ghiboub, M. *et al.* The Epigenetic Reader Protein SP140 Regulates Dendritic Cell Activation, Maturation and Tolerogenic Potential. *Curr. Issues Mol. Biol.* **45**, 4228–4245 (2023).

67. Kotov, D. I. *et al.* Early cellular mechanisms of type I interferon-driven susceptibility to tuberculosis. *Cell* **186**, 5536-5553.e22 (2023).

68. Bottomley, M. J. *et al.* The SAND domain structure defines a novel DNA-binding fold in transcriptional regulation. *Nat. Struct. Biol.* **8**, 626–633 (2001).

69. Bienz, M. The PHD finger, a nuclear protein-interaction domain. *Trends Biochem. Sci.* **31**, 35–40 (2006).

70. Zucchelli, C. *et al.* Sp140 is a multi-SUMO-1 target and its PHD finger promotes SUMOylation of the adjacent Bromodomain. *Biochim. Biophys. Acta Gen. Subj.* **1863**, 456–465 (2019).

71. Filippakopoulos, P. *et al.* Histone recognition and large-scale structural analysis of the human bromodomain family. *Cell* **149**, 214–231 (2012).

72. Huoh, Y.-S. *et al.* Dual functions of Aire CARD multimerization in the transcriptional regulation of T cell tolerance. *Nat. Commun.* **11**, 1625 (2020).

73. Gao, A. *et al.* Evolution of weak cooperative interactions for biological specificity. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E11053–E11060 (2018).

74. Shin, Y. & Brangwynne, C. P. Liquid phase condensation in cell physiology and disease. *Science* **357**, (2017).

75. Su, X. *et al.* Phase separation of signaling molecules promotes T cell receptor signal transduction. *Science* **352**, 595–599 (2016).

76. Sabari, B. R. *et al.* Coactivator condensation at super-enhancers links phase separation and gene control. *Science* **361**, (2018).

77. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–124 (2012).

78. Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.* **42**, 1118–1125 (2010).

79. International Multiple Sclerosis Genetics Consortium (IMSGC) *et al.* Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nat. Genet.* **45**, 1353–1360 (2013).

80. Di Bernardo, M. C. *et al.* A genome-wide association study identifies six susceptibility loci for chronic lymphocytic leukemia. *Nat. Genet.* **40**, 1204–1210 (2008).

81. Sillé, F. C. M., Thomas, R., Smith, M. T., Conde, L. & Skibola, C. F. Post-GWAS functional characterization of susceptibility variants for chronic lymphocytic leukemia. *PLoS One* **7**, e29632 (2012).

82. Matesanz, F. *et al.* A functional variant that affects exon-skipping and protein expression of SP140 as genetic mechanism predisposing to multiple sclerosis. *Hum. Mol. Genet.* **24**, 5619–5627 (2015).

83. Tanagala, K. K. K. *et al.* SP140 inhibits STAT1 signaling, induces IFN-γ in tumor-associated macrophages, and is a predictive biomarker of immunotherapy response. *J Immunother Cancer* **10**, (2022).

84. Chen, D. S. & Mellman, I. Oncology meets immunology: the cancer-immunity cycle. *Immunity* **39**, 1–10 (2013).

85. Yu, B. *et al.* A CIC-related-epigenetic factors-based model associated with prediction, the tumor microenvironment and drug sensitivity in osteosarcoma. *Sci. Rep.* **14**, 1308 (2024).

86. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory activities of transposable elements: from conflicts to benefits. *Nat. Rev. Genet.* **18**, 71–86 (2017).

87. Bourque, G. *et al.* Ten things you should know about transposable elements. *Genome Biol.* **19**, 199 (2018).

88. Kovaka, S. *et al.* Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 278 (2019).

89. Wilks, C., Gaddipati, P., Nellore, A. & Langmead, B. Snaptron: querying splicing patterns across tens of thousands of RNA-seq samples. *Bioinformatics* **34**, 114–116 (2018).

90. Ghandi, M. *et al.* Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569**, 503–508 (2019).

91. Li, Q. *et al.* RJunBase: a database of RNA splice junctions in human normal and cancerous tissues. *Nucleic Acids Res.* **49**, D201–D211 (2021).

92. Kelly, C. J., Chitko-McKown, C. G. & Chuong, E. B. Ruminant-specific retrotransposons shape regulatory evolution of bovine immunity. *Genome Res.* **32**, 1474–1486 (2022).

93. Horton, I., Kelly, C. J., Dziulko, A., Simpson, D. M. & Chuong, E. B. Mouse B2 SINE elements function as IFN-inducible enhancers. *Elife* **12**, (2023).

94. Layer, R. M. *et al.* GIGGLE: a search engine for large-scale integrated genome analysis. *Nat. Methods* **15**, 123–126 (2018).

95. Frost, J. M. *et al.* Regulation of human trophoblast gene expression by endogenous retroviruses. *Nat. Struct. Mol. Biol.* **30**, 527–538 (2023).

96. Du, C. *et al.* Regulation of endogenous retrovirus-derived regulatory elements by GATA2/3 and MSX2 in human trophoblast stem cells. *Genome Res.* **33**, 197–207 (2023).

97. Gao, L. *et al.* Single-cell analysis reveals transcriptomic and epigenomic impacts on the maternal-fetal interface following SARS-CoV-2 infection. *Nat. Cell Biol.* **25**, 1047–1060 (2023).

98. Yu, M. *et al.* Endogenous retrovirus-derived enhancers confer the transcriptional regulation of human trophoblast syncytialization. *Nucleic Acids Res.* **51**, 4745–4759 (2023).

99. Bailey, T. L., Johnson, J., Grant, C. E. & Noble, W. S. The MEME Suite. *Nucleic Acids Res.* **43**, W39-49 (2015).

100. Zheng, R. *et al.* Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. *Nucleic Acids Res.* **47**, D729–D735 (2019).

101. Wagner, E. F. & Nebreda, A. R. Signal integration by JNK and p38 MAPK pathways in cancer development. *Nat. Rev. Cancer* **9**, 537–549 (2009).

102. Saare, M. *et al.* SP140L, an Evolutionarily Recent Member of the SP100 Family, Is an Autoantigen in Primary Biliary Cirrhosis. *J Immunol Res* **2015**, 526518 (2015).

103. Yeo, N. C. *et al.* An enhanced CRISPR repressor for targeted mammalian gene regulation. *Nat. Methods* **15**, 611–616 (2018).

104. Liu, T., Zhang, L., Joo, D. & Sun, S.-C. NF-κB signaling in inflammation. *Signal Transduct. Target. Ther.* **2**, 17023 (2017).

105. Park, M. H. & Hong, J. T. Roles of NF-κB in cancer and inflammatory diseases and their therapeutic approaches. *Cells* **5**, 15 (2016).

106. Jang, H. S. *et al.* Author Correction: Transposable elements drive widespread expression of oncogenes in human cancers. *Nat. Genet.* **51**, 920 (2019).

107. Deniz, Ö. *et al.* Endogenous retroviruses are a source of enhancers with oncogenic potential in acute myeloid leukaemia. *Nat. Commun.* **11**, 3506 (2020).

108. Karttunen, K. *et al.* Transposable elements as tissue-specific enhancers in cancers of endodermal lineage. *Nat. Commun.* **14**, 5313 (2023).

109. Grillo, G. *et al.* Transposable Elements Are Co-opted as Oncogenic Regulatory Elements by Lineage-Specific Transcription Factors in Prostate Cancer. *Cancer Discov.* **13**, 2470–2487 (2023).

110. Matthews, C. P., Colburn, N. H. & Young, M. R. AP-1 a target for cancer prevention. *Curr. Cancer Drug Targets* **7**, 317–324 (2007).

111. Dhillon, A. S., Hagan, S., Rath, O. & Kolch, W. MAP kinase signalling pathways in cancer. *Oncogene* **26**, 3279–3290 (2007).

112. Tasdemir, N. *et al.* BRD4 Connects Enhancer Remodeling to Senescence Immune Surveillance. *Cancer Discov.* **6**, 612–629 (2016).

113. Deininger, P. L., Batzer, M. A., Hutchison, C. A., 3rd & Edgell, M. H. Master genes in mammalian repetitive DNA amplification. *Trends Genet.* **8**, 307–311 (1992).

114. Shen, M. R., Batzer, M. A. & Deininger, P. L. Evolution of the master Alu gene(s). *J. Mol. Evol.* **33**, 311–320 (1991).

115. Martínez-Barricarte, R. *et al.* Human C3 mutation reveals a mechanism of dense deposit disease pathogenesis and provides insights into complement activation and regulation. *J. Clin. Invest.* **120**, 3702–3712 (2010).

116. Chan, B. C. L., Lam, C. W. K., Tam, L.-S. & Wong, C. K. IL33: Roles in allergic inflammation and therapeutic perspectives. *Front. Immunol.* **10**, 364 (2019).

117. Schoeps, B., Frädrich, J. & Krüger, A. Cut loose TIMP-1: an emerging cytokine in inflammation. *Trends Cell Biol.* **33**, 413–426 (2023).

118. Stevens, E. A., Mezrich, J. D. & Bradfield, C. A. The aryl hydrocarbon receptor: a perspective on potential roles in the immune system. *Immunology* **127**, 299–311 (2009).

119. Gaidt, M. M. *et al.* Self-guarding of MORC3 enables virulence factor-triggered immunity. *Nature* **600**, 138–142 (2021).

120. Assouvie, A. *et al.* A genetic variant controls interferon-β gene expression in human myeloid cells by preventing C/EBP-β binding on a conserved enhancer. *PLoS Genet.* **16**, e1009090 (2020).

121. Ferri, F. *et al.* TRIM33 switches off Ifnb1 gene transcription during the late phase of macrophage activation. *Nat. Commun.* **6**, 8900 (2015).

122. Decque, A. *et al.* Sumoylation coordinates the repression of inflammatory and anti-viral gene-expression programs during innate sensing. *Nat. Immunol.* **17**, 140–149 (2016).

123. Frankish, A. *et al.* GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Res.* **51**, D942–D949 (2023).

124. Skene, P. J. & Henikoff, S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *Elife* **6**, (2017).

125. Nassar, L. R. *et al.* The UCSC Genome Browser database: 2023 update. *Nucleic Acids Res.* **51**, D1188–D1195 (2023).

126. Bouchier-Hayes, L. & Martin, S. J. CARD games in apoptosis and immunity. *EMBO Rep.* **3**, 616–621 (2002).

127. Damiano, J. S. & Reed, J. C. CARD proteins as therapeutic targets in cancer. *Curr. Drug Targets* **5**, 367–374 (2004).

128. Park, H. H. Caspase recruitment domains for protein interactions in cellular signaling (Review). *Int. J. Mol. Med.* **43**, 1119–1127 (2019).