TITLE PAGE

Title: Endogenous retroviruses mediate transcriptional rewiring in response to oncogenic signaling in colorectal cancer

Authors: Atma Ivancevic, David M. Simpson, Edward B. Chuong

Corresponding Author:
Edward B. Chuong
edward.chuong@colorado.edu
596 UCB
Boulder, CO 80309
303-735-8573

Affiliations:
BioFrontiers Institute and Department of Molecular, Cellular & Developmental Biology, University of Colorado Boulder

Keywords: Endogenous retroviruses, epigenomics, colorectal cancer, AP1 signaling

Running title: ERV-derived enhancers in colorectal cancer

Abstract: Cancer cells exhibit rewired transcriptional regulatory networks that promote tumor growth and survival. However, the processes that configure these pathological networks remain poorly understood. Through a pan-cancer epigenomic analysis, we found that primate-specific endogenous retroviruses (ERVs) are an abundant source of enhancers that show cancer-specific activity. In colorectal cancer and other epithelial tumors, oncogenic MAPK/AP1 signaling drives activation of enhancers derived from the primate-specific ERV family LTR10. Functional studies in colorectal cancer cells revealed that LTR10 elements are involved in regulating multiple genes associated with tumorigenesis, including *ATG12*. Within the human population, individual LTR10 elements show frequent structural variation resulting from a highly mutable internal tandem repeat region, which affects AP1 binding activity. Our findings reveal that ERV-derived enhancers mediate transcriptional dysregulation in response to oncogenic signaling, and shape the evolution of cancer-specific regulatory networks.

Main Text:

INTRODUCTION

Cancer cells undergo global transcriptional changes resulting from genetic and epigenetic alterations during tumorigenesis (You and Jones 2012). While regulatory remodeling can arise from somatic non-coding mutations (Rheinbay et al. 2020), epigenomic studies have revealed that transformation is associated with aberrant epigenetic activation of enhancer sequences that are typically silenced in normal tissues (Cohen et al. 2017; Chapuy et al. 2013; Roe et al. 2017). Pathological enhancer activity is an established mechanism underlying tumorigenesis and therapy resistance, and therapeutic modulation of enhancer activity is an active area of investigation (Hnisz et al. 2015; Bradner, Hnisz, and Young 2017; Sur and Taipale 2016; Flavahan, Gaskell, and Bernstein 2017). However, we have a limited understanding of the molecular processes that shape and establish the enhancer landscapes of cancer cells.

Transposable elements (TEs) including endogenous retroviruses (ERVs) represent a potentially significant source of enhancers that could shape cancer-specific gene regulation (Jansz and Faulkner 2021). Many cancers exhibit genome-wide transcriptional reactivation of TEs, which can directly impact cells by promoting oncogenic mutations and stimulating immune signaling (Kong et al. 2019; Shukla et al. 2013; Rodriguez-Martin et al. 2020; Rodić et al. 2014). In addition, the reactivation of TEs is increasingly recognized to have gene regulatory consequences in cancer cells (Artem Babaian and Mager 2016). Several transcriptomic studies have uncovered TEs as a source of cancer-specific alternative promoters across many types of cancer, with some examples shown to drive oncogene expression (A. Babaian et al. 2016; Lamprecht et al. 2010; Edginton-White et al. 2019; Jang et al. 2019). TEs also show chromatin signatures of enhancer activity in cancer cell lines (Wang et al. 2007; Sundaram et al. 2014; Jacques, Jeyakani, and Bourque 2013), yet their functional relevance in patient tumors has remained largely unexplored. A recent study identified and characterized ERV-derived enhancers with oncogenic effects in acute myeloid leukemia (Deniz et al. 2020), but the prevalence and mechanisms of TE-derived enhancer activity are unknown for most cancer types.

Here, we leveraged the availability of published cancer epigenome datasets to understand how TEs influence enhancer landscapes and gene regulation across cancer types. Our pan-cancer analysis revealed that elements from a primate-specific ERV named LTR10 show enhancer activity in many epithelial tumors, and this activity is regulated by MAPK/AP1 signaling. We conducted functional studies in HCT116 colorectal cancer cells, and found that LTR10 elements regulate AP1-dependent gene expression at multiple loci that include genes with established roles in tumorigenesis. Finally, we discovered that LTR10 elements contain highly mutable sequences that potentially contribute genomic variation affecting cancer-specific gene expression. Our work implicates ERVs as a source of pathological regulatory variants that facilitate transcriptional rewiring in cancer.

RESULTS

To assess the contribution of TEs to cancer cell epigenomes, we analyzed aggregate chromatin accessibility maps from 21 human cancers generated by The Cancer Genome Atlas project (Corces et al. 2018). We defined cancer-specific subsets of accessible regions by removing regions that show evidence of regulatory activity in any healthy adult tissue profiled by the Roadmap Consortium (Fig 1A, Methods) (Roadmap Epigenomics Consortium et al. 2015). Out of 1315 total repeat families annotated in the human genome, we found 23 families that showed significant enrichment within the accessible chromatin in at least one cancer type (Fig 1B), of which 19 correspond to long terminal repeat (LTR) regions of primate-specific ERVs (Supp Table 1). These observations from chromatin accessibility data generated from primary tumors confirm previous reports of LTR-derived regulatory activity in cancer cell lines (Jacques, Jeyakani, and Bourque 2013; Wang et al. 2007; Deniz et al. 2020), and support a role for ERVs in shaping patient tumor epigenomes.

LTR10 elements exhibit cancer-specific regulatory activity

To further investigate the cancer-specific regulatory activity of ERVs, we focused on LTR10 elements, which were enriched within cancer-specific accessible chromatin for several

types of epithelial tumors including colorectal, stomach, prostate, and lung tumors (Fig 1C, Supp Fig S1A). LTR10 elements (including LTR10A-G, n=2331) are derived from the LTR of the gammaretrovirus HERV-I, which integrated into the anthropoid genome 30 million years ago (Fig 1D, 1E) (Wang et al. 2007). As our initial TCGA analysis was conducted using aggregate data for each tumor type, we first confirmed that LTR10 elements showed recurrent chromatin accessibility across colorectal tumors from multiple individual patients (Fig 1F, Supp Fig S1). We then analyzed epigenomic datasets from the HCT116 colorectal cancer cell line (Akhtar-Zaidi et al. 2012; Baranello et al. 2016; Cohen et al. 2017; Zheng et al. 2019) and found that LTR10A and LTR10F elements exhibit canonical chromatin hallmarks of enhancer activity, including enrichment of histone modifications H3K27ac and H3K4me1, the transcriptional coactivator p300, and RNA Polymerase II occupancy (Fig 1G). We did not observe enhancer-like profiles at LTR10C elements, which have previously been identified as a source of p53 binding sites (Grow et al. 2021; Wang et al. 2007) (Supp Fig S1C). While most LTR10A and LTR10F elements are not transcribed, some show evidence of transcription as promoters for full-length non-coding HERV-I insertions or cellular transcripts (Supp Fig S1D). Therefore, elements derived from the LTR10A and LTR10F subfamilies (hereafter referred to as LTR10 elements) show robust epigenomic signatures associated with enhancer activity in colorectal cancer cells.

We next compared the epigenetic status of LTR10 elements between colorectal cancer cells and normal tissues. In multiple patient-matched epigenomic datasets (Orouji et al. 2021; Bujold et al. 2016), LTR10 elements show globally increased levels of enhancer-associated histone modifications H3K27ac and H3K4me1 in tumor samples compared to adjacent normal colorectal tissues (Fig 2A, 2B). In additional datasets generated from healthy colorectal tissue samples (Cohen et al. 2017; Bernstein et al. 2010; ENCODE Project Consortium 2012; Lister et al. 2009), LTR10 elements do not show any evidence of enhancer activity (Supp Fig S2A). Although we did not examine rare or transient normal cell populations, these observations suggest that LTR10 regulatory activity is generally restricted to tumor cells.

We expanded our analysis to include epigenomic states from all adult tissues (Roadmap Epigenomics Consortium et al. 2015). We found no evidence for LTR10 enhancer activity in normal tissues, but instead observed general enrichment of H3K9me3-associated heterochromatin marks (Fig 2C, Supp Figure S2B, Supp Table 2). To identify factors that directly bind to and potentially repress LTR10 elements, we analyzed the Cistrome database (Zheng et al. 2019) of published human ChIP-Seq datasets to identify transcriptional repressors with evidence for enriched binding within LTR10 elements. Considering all cell types, we found that LTR10 elements are significantly enriched for binding by ZNF562, TRIM28, and SETDB1 (Fig 2D, 2E, Supp Table 3), which are components of the KRAB-ZNF transposon silencing pathway (Imbeault, Helleboid, and Trono 2017). Our analysis suggests that, as expected for most primate-specific TEs (Bruno, Mahgoub, and Macfarlan 2019), LTR10 elements are normally subject to H3K9me3-mediated epigenetic silencing in somatic tissues.

LTR10-derived enhancers are regulated by AP1

To identify which pathways are responsible for cancer-specific reactivation of LTR10 elements, we focused our Cistrome enrichment analysis on activating transcription factors in colorectal cancer cell lines. LTR10 elements were significantly enriched for binding by AP1 complex members (Fig 2F, 2G, Supp Table 3) including the JUND, FOSL1, and ATF3 transcription factors. The LTR10A and LTR10F consensus sequences harbor multiple predicted AP1 binding motifs, which are enriched within LTR10 elements marked by H3K27ac in HCT116 cells (Fig 2H). Moreover, the AP1 motifs are largely absent in other LTR10 subfamilies (Fig 2I). These analyses indicate that the cancer-specific enhancer activity of LTR10 elements is likely driven by sequence-specific recruitment of the AP1 complex.

Dysregulation of AP1 signaling occurs in many cancers, driven by mutations that cause oncogenic activation of the MAPK signaling pathway (Wagner and Nebreda 2009). We first used luciferase reporter assays to test whether LTR10 enhancer activity is affected by modulation of the MAPK/AP1 signaling pathway. We synthesized the LTR10A and LTR10F

consensus sequences as well as variants where the AP1 motifs were disrupted, and cloned the sequences into an enhancer reporter construct. We measured reporter activity in HCT116 cells that were treated for 24 hrs with either TNFa to stimulate signaling or cobimetinib (a MEK1 inhibitor) to inhibit signaling. Consistent with regulation by AP1, cobimetinib treatment caused a decrease in LTR10-driven reporter activity, and TNFa caused an increase (Fig 3A). Overall regulatory activity was greatly reduced in sequences where the AP1 motif was disrupted (Fig 3A). These results show that LTR10 enhancer activity can be directly regulated by modulation of the MAPK/AP1 signaling pathway in cancer cells.

To test the role of the AP1 complex in endogenous LTR10 regulation, we silenced the AP1 component *FOSL1* using CRISPRi to determine the impact on LTR10 transcriptional activity. Using HCT116 cells expressing dCas9-KRAB-MeCP2 (Yeo et al. 2018), we transfected a guide RNA (gRNA) targeting the *FOSL1* transcription start site (TSS) and used RNA-seq to compare gene and TE expression to control cells transfected with a non-targeting gRNA. We first confirmed silencing of *FOSL1* (Supp Fig S3A, S3B), then analyzed TE transcript expression summarized at the family level to account for reads mapping to multiple insertions of the same TE (Jin et al. 2015). This analysis revealed that full-length LTR10/HERV-I elements were significantly downregulated upon silencing *FOSL1* (Fig 3B), supporting a direct role for the AP1 complex in regulating LTR10 activity.

Next, we investigated how endogenous LTR10 elements respond to modulation of MAPK/AP1 signaling at the chromatin level. We treated HCT116 cells with either cobimetinib or TNFa for 24 hrs and profiled each response using RNA-seq and H3K27ac CUT&RUN (Supp Fig S3C, S3D, S3E, S3F). Consistent with our reporter assay results, our RNA-seq analysis showed that full-length LTR10/HERV-I transcripts were significantly downregulated upon cobimetinib treatment, and upregulated upon TNFa treatment (Fig 3B). LTR10 elements showed similar responses based on H3K27ac CUT&RUN signal, exhibiting significant enrichment within the genome-wide set of predicted enhancers downregulated by cobimetinib or upregulated by TNFa (Fig 3C, 3D). We also observed

clear TNFa-induced H3K27ac signal over LTR10 elements in a published dataset of SW480 colorectal cancer cells (Rahnamoun et al. 2017) (Supp Fig S3G). These results indicate that LTR10 elements represent a significant subset of genome-wide enhancers and transcripts in HCT116 cells that are directly modulated by MAPK/AP1 signaling.

LTR10 elements regulate cancer-specific pathological gene expression

To determine whether any LTR10-derived enhancers have a functional effect on MAPK-dependent gene expression in colorectal cancer cells, we used our RNA-seq and CUT&RUN data from HCT116 cells to identify elements predicted to have gene regulatory activity. We defined a set of 620 MAPK-regulated genes that were upregulated by TNFa and downregulated by cobimetinib (Fig 3E, Supp Table 4). We identified LTR10 elements predicted to regulate these genes using the activity by contact model (Fulco et al. 2019) to assign enhancer-gene targets based on H3K27ac signal and chromatin interaction data (Supp Table 5). This yielded 57 LTR10-derived enhancers predicted to regulate 74 MAPK-regulated genes, including many with established roles in cancer pathophysiology (Fig 3F, 3G, Supp Table 6).

We next used CRISPRi to experimentally silence predicted LTR10-derived enhancers in HCT116 cells. We first focused on a predicted tumor-specific enhancer (LTR10.ATG12) formed by two LTR10F elements on chromosome 5, located 87 kb from predicted target genes *ATG12* and *AP3S1* (Fig 4A). The LTR10.ATG12 element shows characteristic epigenomic signatures of tumor-specific enhancer activity and binding by AP1 transcription factors. To silence the element, we stably transfected cells with a pool of four gRNAs targeting unique sequences within the element. As a positive control, we separately transfected cells with a single guide RNA targeting the *ATG12* TSS. We determined the genome-wide transcriptional effects of CRISPRi silencing by using RNA-seq to compare gene expression against cells transfected with a non-targeting gRNA.

We verified that silencing *ATG12* resulted in the expected specific downregulation of *ATG12* (Supp Fig S4A, S4B). In cells where the LTR10.ATG12 element was silenced, the

most significantly downregulated genes were *ATG12* and *AP3S1* (Fig 4B, 4C, Supp Table 7), confirming the predicted enhancer function of the element. In addition, we observed downregulation of other nearby genes including *ARL14EPL* (Fig 4B, 4C), consistent with enhancer activity that affects multiple genes in the locus. Genome-wide, we observed differential regulation of other genes, possibly due to indirect effects from target gene knockdown or off-target silencing of other LTR10 elements (Supp Fig 4C). Notably, we observed that multiple genes regulated by LTR10.ATG12 showed similar patterns of transcriptional downregulation in response to *FOSL* silencing and cobimetinib treatment (Fig 4D). These results indicate that LTR10.ATG12 acts as an enhancer that controls AP1-dependent transcriptional activation of multiple genes in the *ATG12/AP3S1* locus in HCT116 cells.

The *ATG12* gene encodes a ubiquitin-like modifier required for macroautophagy as well as mitochondrial homeostasis and apoptosis (Haller et al. 2014; Rubinstein et al. 2011; Radoshevich et al. 2010; Hanada et al. 2007). Expression of ATG12 is associated with tumorigenesis and therapy resistance in colorectal and gastric cancer (Hu et al. 2018; YiRen et al. 2017), but the mechanism of cancer-specific regulation of ATG12 has not been characterized. Therefore, we aimed to determine whether the LTR10.ATG12 enhancer was responsible for regulating ATG12 expression and activity in HCT116 cells. First, we validated that silencing the enhancer resulted in decreased ATG12 protein levels by immunoblotting (Fig 4E). In cells where either *ATG12* or the enhancer was silenced, there was a clear reduction in protein levels of both free ATG12 and the ATG3-ATG12 conjugate. There was minimal knockdown effect on the levels of the ATG5-ATG12 conjugate, which has previously been observed in *ATG12* silencing experiments and is due to the high stability of the ATG5-ATG12 complex (Haller et al. 2014).

Next, we tested whether ATG12-dependent functions require the activity of the LTR10.ATG12 enhancer. We treated each cell line with staurosporine (STS) to trigger mitochondrial apoptosis, which is dependent on free ATG12 binding to Bcl-2 (Rubinstein et al. 2011). In cells where either *ATG12* or the enhancer was silenced, we observed

significantly reduced caspase 3/7 activity, indicating defective mitochondrial apoptosis (Fig 4F). We did not detect differences in macroautophagy in cells treated with bafilomycin (Supp Fig S4D), consistent with the lack of knockdown of the ATG5-ATG12 conjugate (Hanada et al. 2007). Our experimental results from silencing both *ATG12* and the enhancer are concordant with previous studies directly silencing *ATG12* using siRNAs in other cancer cell lines (Rubinstein et al. 2011; Haller et al. 2014). Together, these experiments demonstrate that the LTR10.ATG12 enhancer is functionally important for ATG12-dependent activity in HCT116 cells.

Having established the regulatory activity of LTR10.ATG12, we selected 5 additional LTR10 elements for characterization based on their proximity to predicted target genes with established cancer relevance (Fig 5). We separately silenced each LTR10 element using CRISPRi and selected one element (LTR10.KDM6A) to delete using CRISPR/Cas9, due to its intronic location (Supp Fig S5). We used RNA-seq to determine the transcriptional consequences of perturbing each element. As we observed with LTR10.ATG12, we observed local downregulation of multiple genes within 1.5 MB of each targeted element, confirming their activity as functional enhancers in HCT116 cells. These included *XRCC4, TMEM167A, VCAN, NES, FGF2, AGPAT5, MAOB,* and *MIR222HG* (Fig 5). For three elements (LTR10.MEF2D, LTR10.MCPH1 and LTR10.KDM6A), the predicted proximal target gene did not show significant expression changes, but we instead observed downregulation of other MAPK-regulated genes near the element (Fig 5G, 5I, 5J). Collectively, our characterization of six LTR10 elements identified 21 genes that are regulated by LTR10 elements, most (17/21) of which are regulated by MAPK signaling based on our RNA-seq data. These experiments suggest that LTR10-derived enhancer elements may exert a genome-wide impact on MAPK-dependent gene expression in cancer cells.

LTR10 elements contain highly mutable VNTRs

Lastly, we investigated variation at LTR10 elements across 15,708 human genomes profiled by the Genome Aggregation database (gnomAD) (Karczewski et al. 2020). All

LTR10 insertions are fixed, but we observed an unexpected enrichment of >10bp indel structural variants affecting the AP1 motif region specific to LTR10A and LTR10F, but not other LTR10 subfamilies such as LTR10C (Fig 6A). Further sequence inspection revealed that LTR10A and LTR10F elements contain an internal variable number of tandem repeats (VNTR) region, composed of a 28-30 bp sequence that includes the AP1 motif (Fig 6B, Supp Fig S6A). Individual elements annotated in the reference genome show extensive variation in tandem repeat length, with up to 33 copies of the AP1 motif (Supp Fig S6B). The number of motifs strongly correlated with H3K27ac and FOSL1 binding activity in HCT116 cells (Supp Fig S6B), suggesting that tandem repeat length affects AP1-dependent regulation of individual elements. Across the human population, LTR10A and LTR10F elements harbor many rare and common indel structural variants of lengths that follow a 28-30 bp periodicity, and this pattern is absent in LTR10C elements which lack the tandem repeat region (Fig 6C, 6D). These elevated levels of polymorphism across copies and individuals are characteristic of unstable tandem repeat regions (Gymrek 2017), and suggest that LTR10 VNTR regions may be a common source of genomic regulatory variation.

Accurately genotyping tandem repeat length polymorphisms remains a major challenge using short-read data, therefore we validated the presence of LTR10 VNTR polymorphisms using long-read structural variant calls generated from 15 individuals (Audano et al. 2019). We recovered indel structural variants within 24 distinct LTR10A and LTR10F elements, which also showed 28-30 bp periodicity (Fig 6D, Supp Fig S6C). We confirmed the presence of additional LTR10 VNTR indels using a separate long-read dataset from 25 Asian individuals (Fig 6E, Supp Fig S6D-I) (Quan et al. 2021). At the LTR10.ATG12 locus, we observed multiple indels supported by both short-read and long-read data that are predicted to affect AP1 motif copy number (Fig 6E, Supp Fig S6E). At a genome-wide level, LTR10 elements were a significantly enriched source of long-read indels, despite being fixed in the population (Fig 6F). Finally, in addition to finding extensive germline genetic variation affecting LTR10 elements, we also found evidence for tumor-specific somatic mutations within LTR10 elements, based on tumor-specific variants

called by the International Cancer Genome Consortium (Supp Fig S6J) (ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium 2020). Together, these analyses reveal that LTR10 VNTR regions are a source of non-coding polymorphisms, which potentially contribute to both germline and somatic regulatory variation in cancer cells.

DISCUSSION

Our findings build on a growing body of evidence implicating TEs as an important source of pathological regulatory activity in cancer (Jang et al. 2019; Deniz et al. 2020; Lamprecht et al. 2010; A. Babaian et al. 2016). Although TEs have been widely documented to show elevated transcriptional activity in cancer cells, the mechanisms underlying the regulatory reactivation of TEs in cancer have remained poorly understood. Our study demonstrates that oncogenic MAPK/AP1 signaling drives regulatory activation of LTR10 and other ERVs in epithelial tumors, which act as enhancers that mediate AP1-dependent transcriptional dysregulation. While our functional experiments were conducted in a cell line model, our analysis indicates that LTR10 elements show epigenomic signatures of enhancer activity in many patient tumors. The specific consequences of ERV-derived enhancer activity in cancer cells are likely to vary depending on genomic context and tumor type, and individual elements may exert a spectrum of pathological, anti-tumorigenic, or nonfunctional regulatory effects. Nonetheless, our findings demonstrate how ERV-derived enhancers act as a regulatory link between oncogenic signaling and genome-wide transcriptional dysregulation in cancer cells.

Our study used CRISPR to functionally validate multiple LTR10 elements with predicted enhancer activity, revealing their involvement in regulating genes with cancer relevance. As a case example, we demonstrated that a LTR10 element was required for ATG12-dependent mitochondrial apoptosis in HCT116 colorectal cancer cells. ATG12 expression has been implicated in tumorigenesis and therapy resistance in several cancer types, making it a candidate for therapeutic targeting (Rubinsztein, Codogno, and Levine 2012; Hu et al. 2018). Notably, MAPK inhibitor treatment results in silencing LTR10

regulatory activity, suggesting that ERV enhancer inactivation may be an underlying mechanism contributing to therapeutic MAPK inhibition.

Finally, we discovered that LTR10 elements are frequently affected by indels that could influence their regulatory activity. Although all LTR10 insertions are fixed in the human population, they contain internal tandem repeats that show high levels of length polymorphism associated with repeat instability. Germline or somatic variation in AP1 motif copy number within these elements may alter cancer-specific enhancer landscapes, and we speculate that LTR10 VNTRs may be subject to further mutation in cancer cells with microsatellite instability (Hung et al. 2019). Our study of LTR10 highlights how TEs that are normally silenced can become reactivated in cancer and cause aberrant gene expression. For elements that promote pathogenesis, their restricted activity in age-associated diseases like cancer may result in reduced or nearly neutral fitness consequences. Therefore, the accumulation of TEs subject to epigenetic silencing may be a fundamental process shaping the evolution of cancer-specific gene regulatory networks.

METHODS

Cell culture

The HCT116 cell line was purchased from ATCC and cultured in McCoy's 5A medium supplemented with 10% FBS and 1% penicillin/streptomycin (Gibco). Cells were cultured at 37°C in 5% carbon dioxide. Transfections were performed using FuGENE (Promega). For treatments modulating MAPK signaling, HCT116 cells were untreated or treated for 24 hrs with 1 uM Cobimetinib, 100 ng/mL TNF alpha, or DMSO.

CRISPR-mediated silencing and knockout of LTR10s

For CRISPR-mediated silencing (e.g., CRISPRi) of select LTR10 elements and gene transcription start sites (TSS), a HCT116 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 the Super PiggyBac transposase expression vector (SPBT) into HCT116 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 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). Chemically competent Stable E. Coli (New England Biolabs) was 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 HCT116 dCas9-KRAB-MeCP2 polyclonal stable line. The following number of uniquely-mapping gRNA plasmids were designed per target based on the pre-computed UCSC hg38 CRISPR target track: ATG12 (1), GFP (1), FOSL1 (1), LTR10.ATG12 (4), LTR10.FGF2 (2), LTR10.MCPH1 (3), LTR10.MEF2D (2), LTR10.XRCC4 (2). The same total amount of gRNA plasmid was used for transfections involving one or multiple gRNAs. 24 hours post-transfection, cells were treated with Puromycin to select for integration of the sgRNA expression cassette(s). Selection was maintained for 5 days prior to transcriptional analyses.

For CRISPR-mediated knockout of LTR10.KDM6A, 2 gRNAs (1 specific to each flank of the element) were identified and synthesized as sgRNAs by IDT. Using IDT's AltR technology, RNP complexes were generated in vitro, and electroporated into HCT116 cells using the Neon system (ThermoFisher Scientific). Clonal lines were isolated using the array dilution method in a 96-well plate format, and single clones were identified and screened for homozygous deletions by PCR using both flanking and internal primer pairs at the expected deletion site.

Cell autophagy and apoptosis assays

For assaying mitochondrial apoptosis, HCT116 CRISPRi cell lines were treated for 12 hours with Staurosporine (STS) at 0.5 uM or DMSO (vehicle) followed by measurement of Caspase activity via the Caspase-Glo 3/7 assay (Promega). Results are representative of at least 3 independent experiments. For assaying autophagy, HCT116 CRISPRi cell lines were untreated or treated with Bafilomycin A at 10 nM or 100 nM for 6 hrs and 18 hrs, followed by LC3B Western blotting. Results are representative of at least 3 independent experiments.

Western blots

For ATG12 Western blots, cell lysates were prepared with MPER buffer (ThermoFisher Scientific). For LC3B Western blots, cell lysates were prepared with RIPA buffer. All cell Lysates were resuspended in 4x NuPage LDS Sample buffer containing reducing agent (ThermoFisher Scientific). For ATG12 Western blots, total protein was concentrated and size-selected by passing through an Amicon Ultra 10K column (Millipore), retaining the high molecular weight fraction, and 40 ug of protein was loaded per lane. For LC3B Western blots, 2 ug total protein was loaded per lane. Antibodies used were as follows: ATG12 (cat #4180T, Cell Signaling Technologies), Beta-Actin (cat # 3700T, Cell Signaling Technologies), LC3B (cat # NB100-2220, Novus Biologicals). Results are representative of at least 3 independent experiments.

Luciferase assay

Reporter assays were conducted in HCT116 cells using the secreted NanoLuc enhancer activity reporter pNL3.3 (Promega) and normalized against a constitutively active firefly luciferase reporter vector, pGL4.50 (Promega). LTR10 consensus sequences for subfamilies LTR10A and LTR10F were downloaded from Dfam. AP1 motifs within LTR10A and LTR10F were shuffled as follows: LTR10A (first two motifs): cctgagtcacc to cagccccgtta; LTR10A (third motif): cttagtcacc to cagtttaccc; LTR10F (all three motifs): cctgactcatt to cgtatccttac. Sequences are available in Supp Table 8. Due to their high repeat content, consensus sequences were synthesized as multiple fragments (Integrated DNA Technologies, Twist BioScience) and then assembled into pNL3.3 enhancer reporter plasmids using Gibson Assembly (New England Biolabs). Each cloned reporter plasmid was verified by Sanger sequencing (Quintara Bio). To assay reporter activity, HCT116 cells were transfected with a reporter construct as well as the pGL4.50 construct constitutively expressing firefly luciferase. 24 hrs after transfection, media was replaced with media containing 1 uM Cobimetinib, 100 ng/mL TNF alpha, or DMSO (vehicle). 24 hours following treatment, luminescence was measured using the NanoGlo Dual Luciferase Reporter Assay System (Promega). All experiments were performed with 3 treatment replicates per condition in a 96-well plate format, and all reported values were normalized

to firefly co-transfection controls. Results are representative of at least 3 independent experiments. Barplots are presented as mean +/- s.d.

RNA-seq

Sequencing libraries were prepared from RNA harvested from treatment or transfection replicates. Total RNA was extracted using the Quick-RNA Miniprep Plus Kit (Zymo Research). PolyA enrichment and library preparation was performed using the KAPA BioSystems mRNA HyperPrep Kit according to the manufacturer's protocols. Briefly, 500 ng of RNA was used as input, and KAPA BioSystems single-index or unique dual-index adapters were added at a final concentration of 7 nM. Purified, adapter-ligated library was amplified for a total of 11 cycles following the manufacturer's protocol. The final libraries were pooled and sequenced on an Illumina NovaSeq 6000 (University of Colorado Genomics Core) as 150 bp paired-end reads.

CUT&RUN

Libraries were prepared from treatment replicates. Approximately $5x10^5$ viable cells were used for each CUT&RUN reaction, and pulldowns were generated following the protocol from (Meers et al. 2019). All buffers were prepared according to the "High Ca^{2+}/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), rabbit anti-H3K27ac (1:100). pA-MNase (gift from Steven Henikoff) 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, 14x(15 s at 98°C, 10 s at 60°C), 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.

Processing of sequencing data

Reads obtained from our own datasets and from published studies were reprocessed using a uniform analysis pipeline. FASTQ reads were evaluated using FastQC (v0.11.8) and MultiQC (v1.7), then trimmed using BBDuk/BBMap (v38.05). For ATAC-seq, ChIP-seq, and CUT&RUN datasets, reads were aligned to the hg38 human genome using BWA (v0.7.15) and filtered for uniquely mapping reads (MAPQ > 10) with samtools (v1.10). ChIP-Seq and ATAC-seq peak calls and normalized signal coverage bigwig plots were generated using MACS2 (v2.1.1) (with setting --SPMR). CUT&RUN peak calls were generated using MACS2 in paired-end mode using a relaxed p-value threshold without background normalization (--format BAMPE --pvalue 0.01 --SPMR -B --call-summits). MACS2 was also run in single-end mode with additional parameters --shift -75 and --extsize 150, and Bedtools (v2.28.0) was used to merge peaks across the two modes of peak calling for each sample (bedtools merge with options -c 5 -o max).

RNA-seq and PRO-seq reads were aligned to hg38 using hisat2 (v2.1.0) with option --no-softclip and filtered for uniquely mapping reads with samtools for MAPQ > 10. Bigwig tracks were generated using the bamCoverage function of deepTools (v3.0.1), with CPM normalisation (ignoring chrX and chrM) and bin size 1bp. Some datasets from TCGA, ENCODE, Cistrome DB and the CEMT Canadian Epigenomes Project were downloaded as post-processed peaks and bigwig files.

TE colocalization analysis

To determine TE family enrichment within regulatory regions, we used GIGGLE (v0.6.3) (Layer et al. 2018) to generate a genomic interval index of all TE families in the hg38 human genome, based on Dfam v2.0 repeat annotation (n=1315 TE families). Regulatory regions (e.g., ATAC, ChIP-Seq, or CUT&RUN peaks) were queried against the TE interval

index using the GIGGLE search function (-g 3209286105 -s). Results were ranked by GIGGLE enrichment score, which is a composite of the product of −log10(P value) and log2(odds ratio). Significantly enriched TE families were defined as those with at least 25 overlaps between TE copies and a set of peak regions, odds ratio over 10, and GIGGLE score over 100 in at least one cancer type.

Defining cancer-specific regulatory elements

To define cancer-specific regulatory elements, we obtained aggregate ATAC-seq regions associated with each tumor type profiled by The Cancer Genome Atlas (Weinstein et al. 2013), which represent a union of recurrent ATAC-seq regions associated with each tumor type. To filter predicted regulatory regions in healthy adult tissues, we used chromHMM regulatory regions defined by the Roadmap project, using healthy adult tissues from categories 1_TssA, 6_EnhG & 7_Enh (e.g. filtering out placental tissues, embryonic stem cells and trophoblast stem cells due to high numbers of active TEs). Cancer-specific regulatory regions were defined using the subtract function of bedtools (option -A) to subtract Roadmap regulatory regions from each cancer peak set.

Transcription factor motif analyses

Motif analysis of LTR10 elements was performed using MEME suite (v5.4.1) in differential enrichment mode (Bailey et al. 2009). HCT116 CUT&RUN H3K27ac-marked LTR10A/F sequences (n=144) were used as input against a background set of unmarked LTR10A/F sequences (n=561), with default settings other than the number of motif repetitions (Any) and 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 (v5.4.1). FIMO (v5.1.0) was then used to extract motif frequency and hg38 genomic coordinates, with p-value threshold set to 1e-4.

Differential analysis using DESeq2

For RNA-seq samples, gene count tables were generated using featureCounts from the subread (v1.6.2) package with the GENCODE v34 annotation gtf to estimate counts at the

gene level, over each exon (including -p to count fragments instead of reads for paired-end reads; -O to assign reads to their overlapping meta-features; -s 2 to specify reverse-strandedness; -t exon to specify the feature type; -g gene_id to specify the attribute type).

To quantify TE expression at the family level, RNA-seq samples were first re-aligned to hg38 using hisat2 with -k 100 to allow multi-mapping reads and --no-softclip to disable soft-clipping of reads. TEtranscripts (v2.1.4) was then used in multi-mapping mode with the GENCODE v34 annotation gtf and hg38 GENCODE TE gtf to assign count values to both genes and TE elements.

For H3K27ac CUT&RUN samples, the bedtools multicov was used to generate a count table of the number of aligned reads that overlap MACS2-defined peak regions. The top 20,000 peaks were extracted from each sample and merged (using bedtools merge with -d 100) to produce the peak file used as input to bedtools multicov.

All count tables were processed with DEseq2 (v1.32.0). Normalized count values were calculated using the default DEseq2 transformation. R packages ggplot2 (v3.3.2), ggrepel (v0.8.2) and apeglm (v1.8.0) were used to visualize differentially expressed genes or TEs. The same DEseq2 analyses were used to identify differentially enriched peak regions between H3K27ac CUT&RUN samples (e.g. in response to MAPK treatment). Significantly differentially enriched regions were queried against the Giggle index of human repeats to identify over-represented TE families.

Prediction of LTR10 enhancer gene targets

LTR10 elements were initially prioritized for CRISPR silencing or deletion based on enhancer predictions from the Activity-by-Contact (ABC) model (Fulco et al. 2019). Publicly available HCT116 ATACseq (GEO accession GSM3593802) and in-house HCT116 H3K27ac CUT&RUN were used as input to the ABC pipeline, as well as the provided averaged human cell line HiC file. Predicted enhancer regions with an ABC interaction

score over 0.001 were intersected with H3K27ac-marked LTR10A/F elements. Putative LTR10 enhancers were then checked for proximity (e.g. within 1.5Mb) to MAPK-regulated genes (i.e. genes that were significantly affected by MAPK treatments Cobimetinib and TNF-alpha, based on inhouse RNAseq).

Evolutionary analysis of LTR10 sequences

Genomic coordinates of LTR10 elements in the hg38 human genome were obtained from Dfam v2.0, based on RepeatMasker v4.0.6 repeat annotation. The nucleotide sequence of each LTR10 element was extracted using the getfasta function from bedtools (using -name+ to include coordinates in the header and -s for strand specificity). VSEARCH v2.14.1 was used to set a minimum length threshold of 200bp for LTR10 sequences (--sortbylength -minseqlength 200), prior to alignment. MUSCLE (v3.8.1551) was used to align the remaining sequences. Jalview (v2.11.1.4) was used to perform a principal component analysis on pairwise similarity scores derived from the multiple sequence alignment.

LTR10 consensus sequences representing each LTR10 subfamily (A-G) were likewise downloaded from Dfam. Sequences were concatenated into one FASTA file and aligned using MUSCLE. FastTree was used to infer a maximum likelihood phylogeny representing the LTR10 subfamily relationships.

The phylogeny of known primate relationships was obtained from TimeTree (Kumar et al. 2017) and the HERV-I insertion estimate was confirmed based on the presence or absence of LTR10 sequences across mammals, using BLAST (v2.7.1+) (Altschul et al. 1990).

VNTR identification

gnomAD (v3.1) VCF files for each hg38 chromosome were filtered for high-confidence indels (FILTER=PASS) using the query function of bcftools (v1.8) with format parameter -f'%CHROM\t%POS0\t%END\t%ID\t%REF\t%ALT\t%AF\t%TYPE\tFILTER=%FILTER\n'.

The remaining indels were then subsetted by size to retain insertions or deletions between 10 to 300bp in length, and chromosome vcfs were concatenated into one whole genome bed file. Bedtools (v2.28.0) was used to intersect the indel bed file with LTR10 elements from each subfamily, based on Dfam (v2.0) repeat annotation. Indels from long-read datasets were likewise filtered by variant type (INS or DEL) and length (50-300bp), then intersected with LTR10 elements. Deletion length versus allele frequency was plotted for each subfamily, from each separate dataset.

DATA AVAILABILITY

Inhouse high-throughput sequencing data has been deposited in the Gene Expression Omnibus (GEO) with the accession code GSE186619. GSE IDs of public datasets used in this study are listed in the figure legends and GitHub repository. The following large-scale databases were also used: Cistrome DB (downloaded Feb 2019), Roadmap/ENCODE (downloaded Feb 2019), The Cancer Genome Atlas (downloaded Sept 2019), CEMT Canadian Epigenome Project (downloaded July 2020), Dfam 2.0 and gnomAD v3.1.

CODE AVAILABILITY

Source code and workflows are available on GitHub:
https://github.com/atmaivancevic/ERV_cancer_enhancers

COMPETING INTEREST STATEMENT

The authors declare no competing interests.

REFERENCES

Akhtar-Zaidi, Batool, Richard Cowper-Sal-lari, Olivia Corradin, Alina Saiakhova, Cynthia F. Bartels, Dheepa Balasubramanian, Lois Myeroff, et al. 2012. “Epigenomic Enhancer Profiling Defines a Signature of Colon Cancer.” *Science* 336 (6082): 736–39.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. “Basic Local Alignment Search Tool.” *Journal of Molecular Biology* 215 (3): 403–10.

Audano, Peter A., Arvis Sulovari, Tina A. Graves-Lindsay, Stuart Cantsilieris, Melanie Sorensen, Annemarie E. Welch, Max L. Dougherty, et al. 2019. “Characterizing the Major Structural Variant Alleles of the Human Genome.” *Cell* 176 (3): 663–75.e19.

Babaian, A., M. T. Romanish, L. Gagnier, L. Y. Kuo, M. M. Karimi, C. Steidl, and D. L.

Mager. 2016. “Onco-Exaptation of an Endogenous Retroviral LTR Drives IRF5 Expression in Hodgkin Lymphoma.” *Oncogene* 35 (19): 2542–46.
Babaian, Artem, and Dixie L. Mager. 2016. “Endogenous Retroviral Promoter Exaptation in Human Cancer.” *Mobile DNA* 7 (December): 24.
Bailey, Timothy L., Mikael Boden, Fabian A. Buske, Martin Frith, Charles E. Grant, Luca Clementi, Jingyuan Ren, Wilfred W. Li, and William S. Noble. 2009. “MEME SUITE: Tools for Motif Discovery and Searching.” *Nucleic Acids Research* 37 (Web Server issue): W202–8.
Baranello, Laura, Damian Wojtowicz, Kairong Cui, Ballachanda N. Devaiah, Hye-Jung Chung, Ka Yim Chan-Salis, Rajarshi Guha, et al. 2016. “RNA Polymerase II Regulates Topoisomerase 1 Activity to Favor Efficient Transcription.” *Cell* 165 (2): 357–71.
Bernstein, Bradley E., John A. Stamatoyannopoulos, Joseph F. Costello, Bing Ren, Aleksandar Milosavljevic, Alexander Meissner, Manolis Kellis, et al. 2010. “The NIH Roadmap Epigenomics Mapping Consortium.” *Nature Biotechnology* 28 (10): 1045–48.
Bradner, James E., Denes Hnisz, and Richard A. Young. 2017. “Transcriptional Addiction in Cancer.” *Cell* 168 (4): 629–43.
Bruno, Melania, Mohamed Mahgoub, and Todd S. Macfarlan. 2019. “The Arms Race Between KRAB–Zinc Finger Proteins and Endogenous Retroelements and Its Impact on Mammals,” December. https://doi.org/10.1146/annurev-genet-112618-043717.
Bujold, David, David Anderson de Lima Morais, Carol Gauthier, Catherine Côté, Maxime Caron, Tony Kwan, Kuang Chung Chen, et al. 2016. “The International Human Epigenome Consortium Data Portal.” *Cell Systems* 3 (5): 496–99.e2.
Chapuy, Bjoern, Michael R. McKeown, Charles Y. Lin, Stefano Monti, Margaretha G. M. Roemer, Jun Qi, Peter B. Rahl, et al. 2013. “Discovery and Characterization of Super-Enhancer-Associated Dependencies in Diffuse Large B Cell Lymphoma.” *Cancer Cell* 24 (6): 777–90.
Cohen, Andrea J., Alina Saiakhova, Olivia Corradin, Jennifer M. Luppino, Katreya Lovrenert, Cynthia F. Bartels, James J. Morrow, et al. 2017. “Hotspots of Aberrant Enhancer Activity Punctuate the Colorectal Cancer Epigenome.” *Nature Communications* 8 (February): 14400.
Corces, M. Ryan, Jeffrey M. Granja, Shadi Shams, Bryan H. Louie, Jose A. Seoane, Wanding Zhou, Tiago C. Silva, et al. 2018. “The Chromatin Accessibility Landscape of Primary Human Cancers.” *Science* 362 (6413). https://doi.org/10.1126/science.aav1898.
Deniz, Özgen, Mamataz Ahmed, Christopher D. Todd, Ana Rio-Machin, Mark A. Dawson, and Miguel R. Branco. 2020. “Endogenous Retroviruses Are a Source of Enhancers with Oncogenic Potential in Acute Myeloid Leukaemia.” *Nature Communications* 11 (1): 3506.
Edginton-White, Benjamin, Pierre Cauchy, Salam A. Assi, Sylvia Hartmann, Arthur G. Riggs, Stephan Mathas, Peter N. Cockerill, and Constanze Bonifer. 2019. “Global Long Terminal Repeat Activation Participates in Establishing the Unique Gene Expression Programme of Classical Hodgkin Lymphoma.” *Leukemia* 33 (6): 1463–74.
ENCODE Project Consortium. 2012. “An Integrated Encyclopedia of DNA Elements in the Human Genome.” *Nature* 489 (7414): 57–74.

Flavahan, William A., Elizabeth Gaskell, and Bradley E. Bernstein. 2017. “Epigenetic Plasticity and the Hallmarks of Cancer.” *Science* 357 (6348). https://doi.org/10.1126/science.aal2380.

Fulco, Charles P., Joseph Nasser, Thouis R. Jones, Glen Munson, Drew T. Bergman, Vidya Subramanian, Sharon R. Grossman, et al. 2019. “Activity-by-Contact Model of Enhancer-Promoter Regulation from Thousands of CRISPR Perturbations.” *Nature Genetics* 51 (12): 1664–69.

Grow, Edward J., Bradley D. Weaver, Christina M. Smith, Jingtao Guo, Paula Stein, Sean C. Shadle, Peter G. Hendrickson, et al. 2021. “p53 Convergently Activates Dux/DUX4 in Embryonic Stem Cells and in Facioscapulohumeral Muscular Dystrophy Cell Models.” *Nature Genetics* 53 (8): 1207–20.

Gymrek, Melissa. 2017. “A Genomic View of Short Tandem Repeats.” *Current Opinion in Genetics & Development* 44 (June): 9–16.

Haller, Martina, Andreas K. Hock, Evangelos Giampazolias, Andrew Oberst, Douglas R. Green, Jayanta Debnath, Kevin M. Ryan, Karen H. Vousden, and Stephen W. G. Tait. 2014. “Ubiquitination and Proteasomal Degradation of ATG12 Regulates Its Proapoptotic Activity.” *Autophagy* 10 (12): 2269–78.

Hanada, Takao, Nobuo N. Noda, Yoshinori Satomi, Yoshinobu Ichimura, Yuko Fujioka, Toshifumi Takao, Fuyuhiko Inagaki, and Yoshinori Ohsumi. 2007. “The Atg12-Atg5 Conjugate Has a Novel E3-like Activity for Protein Lipidation in Autophagy *.” *The Journal of Biological Chemistry* 282 (52): 37298–302.

Hnisz, Denes, Jurian Schuijers, Charles Y. Lin, Abraham S. Weintraub, Brian J. Abraham, Tong Ihn Lee, James E. Bradner, and Richard A. Young. 2015. “Convergence of Developmental and Oncogenic Signaling Pathways at Transcriptional Super-Enhancers.” *Molecular Cell* 58 (2): 362–70.

Hu, J. L., G. Y. He, X. L. Lan, Z. C. Zeng, J. Guan, Y. Ding, X. L. Qian, W. T. Liao, Y. Q. Ding, and L. Liang. 2018. “Inhibition of ATG12-Mediated Autophagy by miR-214 Enhances Radiosensitivity in Colorectal Cancer.” *Oncogenesis* 7 (2): 16.

Hung, Stevephen, Alina Saiakhova, Zachary J. Faber, Cynthia F. Bartels, Devin Neu, Ian Bayles, Evelyn Ojo, et al. 2019. “Mismatch Repair-Signature Mutations Activate Gene Enhancers across Human Colorectal Cancer Epigenomes.” *eLife* 8 (February). https://doi.org/10.7554/eLife.40760.

ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. 2020. “Pan-Cancer Analysis of Whole Genomes.” *Nature* 578 (7793): 82–93.

Imbeault, Michaël, Pierre-Yves Helleboid, and Didier Trono. 2017. “KRAB Zinc-Finger Proteins Contribute to the Evolution of Gene Regulatory Networks.” *Nature* 543 (7646): 550–54.

Jacques, Pierre-Étienne, Justin Jeyakani, and Guillaume Bourque. 2013. “The Majority of Primate-Specific Regulatory Sequences Are Derived from Transposable Elements.” *PLoS Genetics* 9 (5): e1003504.

Jang, Hyo Sik, Nakul M. Shah, Alan Y. Du, Zea Z. Dailey, Erica C. Pehrsson, Paula M. Godoy, David Zhang, et al. 2019. “Transposable Elements Drive Widespread Expression of Oncogenes in Human Cancers.” *Nature Genetics*. https://doi.org/10.1038/s41588-019-0373-3.

Jansz, Natasha, and Geoffrey J. Faulkner. 2021. “Endogenous Retroviruses in the Origins and Treatment of Cancer.” *Genome Biology* 22 (1): 147.

Jin, Ying, Oliver H. Tam, Eric Paniagua, and Molly Hammell. 2015. “TEtranscripts: A Package for Including Transposable Elements in Differential Expression Analysis of RNA-Seq Datasets.” *Bioinformatics* 31 (22): 3593–99.

Karczewski, Konrad J., Laurent C. Francioli, Grace Tiao, Beryl B. Cummings, Jessica Alföldi, Qingbo Wang, Ryan L. Collins, et al. 2020. “The Mutational Constraint Spectrum Quantified from Variation in 141,456 Humans.” *Nature* 581 (7809): 434–43.

Kong, Yu, Christopher M. Rose, Ashley A. Cass, Alexander G. Williams, Martine Darwish, Steve Lianoglou, Peter M. Haverty, et al. 2019. “Transposable Element Expression in Tumors Is Associated with Immune Infiltration and Increased Antigenicity.” *Nature Communications* 10 (1): 5228.

Kumar, Sudhir, Glen Stecher, Michael Suleski, and S. Blair Hedges. 2017. “TimeTree: A Resource for Timelines, Timetrees, and Divergence Times.” *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/msx116.

Lamprecht, Björn, Korden Walter, Stephan Kreher, Raman Kumar, Michael Hummel, Dido Lenze, Karl Köchert, et al. 2010. “Derepression of an Endogenous Long Terminal Repeat Activates the CSF1R Proto-Oncogene in Human Lymphoma.” *Nature Medicine* 16 (5): 571–79.

Layer, Ryan M., Brent S. Pedersen, Tonya DiSera, Gabor T. Marth, Jason Gertz, and Aaron R. Quinlan. 2018. “GIGGLE: A Search Engine for Large-Scale Integrated Genome Analysis.” *Nature Methods* 15 (2): 123–26.

Lister, Ryan, Mattia Pelizzola, Robert H. Dowen, R. David Hawkins, Gary Hon, Julian Tonti-Filippini, Joseph R. Nery, et al. 2009. “Human DNA Methylomes at Base Resolution Show Widespread Epigenomic Differences.” *Nature* 462 (7271): 315–22.

Meers, Michael P., Terri D. Bryson, Jorja G. Henikoff, and Steven Henikoff. 2019. “Improved CUT&RUN Chromatin Profiling Tools.” *eLife* 8 (June). https://doi.org/10.7554/eLife.46314.

Orouji, Elias, Ayush T. Raman, Anand K. Singh, Alexey Sorokin, Emre Arslan, Archit K. Ghosh, Jonathan Schulz, et al. 2021. “Chromatin State Dynamics Confers Specific Therapeutic Strategies in Enhancer Subtypes of Colorectal Cancer.” *Gut*, May. https://doi.org/10.1136/gutjnl-2020-322835.

Quan, Cheng, Yuanfeng Li, Xinyi Liu, Yahui Wang, Jie Ping, Yiming Lu, and Gangqiao Zhou. 2021. “Characterization of Structural Variation in Tibetans Reveals New Evidence of High-Altitude Adaptation and Introgression.” *Genome Biology* 22 (1): 159.

Radoshevich, Lilliana, Lyndsay Murrow, Nan Chen, Estefania Fernandez, Srirupa Roy, Christopher Fung, and Jayanta Debnath. 2010. “ATG12 Conjugation to ATG3 Regulates Mitochondrial Homeostasis and Cell Death.” *Cell* 142 (4): 590–600.

Rahnamoun, Homa, Hanbin Lu, Sascha H. Duttke, Christopher Benner, Christopher K. Glass, and Shannon M. Lauberth. 2017. “Mutant p53 Shapes the Enhancer Landscape of Cancer Cells in Response to Chronic Immune Signaling.” *Nature Communications* 8 (1): 754.

Rheinbay, Esther, Morten Muhlig Nielsen, Federico Abascal, Jeremiah A. Wala, Ofer Shapira, Grace Tiao, Henrik Hornshøj, et al. 2020. “Analyses of Non-Coding Somatic Drivers in 2,658 Cancer Whole Genomes.” *Nature* 578 (7793): 102–11.

Roadmap Epigenomics Consortium, Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, et al. 2015. “Integrative Analysis of 111 Reference Human Epigenomes.” *Nature* 518 (7539): 317–30.

Rodić, Nemanja, Reema Sharma, Rajni Sharma, John Zampella, Lixin Dai, Martin S. Taylor, Ralph H. Hruban, et al. 2014. “Long Interspersed Element-1 Protein Expression Is a Hallmark of Many Human Cancers.” *The American Journal of Pathology* 184 (5): 1280–86.

Rodriguez-Martin, Bernardo, Eva G. Alvarez, Adrian Baez-Ortega, Jorge Zamora, Fran Supek, Jonas Demeulemeester, Martin Santamarina, et al. 2020. “Pan-Cancer Analysis of Whole Genomes Identifies Driver Rearrangements Promoted by LINE-1 Retrotransposition.” *Nature Genetics* 52 (3): 306–19.

Roe, Jae-Seok, Chang-Il Hwang, Tim D. D. Somerville, Joseph P. Milazzo, Eun Jung Lee, Brandon Da Silva, Laura Maiorino, et al. 2017. “Enhancer Reprogramming Promotes Pancreatic Cancer Metastasis.” *Cell* 170 (5): 875–88.e20.

Rubinstein, Assaf D., Miriam Eisenstein, Yaara Ber, Shani Bialik, and Adi Kimchi. 2011. “The Autophagy Protein Atg12 Associates with Antiapoptotic Bcl-2 Family Members to Promote Mitochondrial Apoptosis.” *Molecular Cell* 44 (5): 698–709.

Rubinsztein, David C., Patrice Codogno, and Beth Levine. 2012. “Autophagy Modulation as a Potential Therapeutic Target for Diverse Diseases.” *Nature Reviews. Drug Discovery* 11 (9): 709–30.

Shukla, Ruchi, Kyle R. Upton, Martin Muñoz-Lopez, Daniel J. Gerhardt, Malcolm E. Fisher, Thu Nguyen, Paul M. Brennan, et al. 2013. “Endogenous Retrotransposition Activates Oncogenic Pathways in Hepatocellular Carcinoma.” *Cell* 153 (1): 101–11.

Sundaram, Vasavi, Yong Cheng, Zhihai Ma, Daofeng Li, Xiaoyun Xing, Peter Edge, Michael P. Snyder, and Ting Wang. 2014. “Widespread Contribution of Transposable Elements to the Innovation of Gene Regulatory Networks.” *Genome Research*. https://doi.org/10.1101/gr.168872.113.

Sur, Inderpreet, and Jussi Taipale. 2016. “The Role of Enhancers in Cancer.” *Nature Reviews. Cancer* 16 (8): 483–93.

Wagner, Erwin F., and Angel R. Nebreda. 2009. “Signal Integration by JNK and p38 MAPK Pathways in Cancer Development.” *Nature Reviews. Cancer* 9 (8): 537–49.

Wang, Ting, Jue Zeng, Craig B. Lowe, Robert G. Sellers, Sofie R. Salama, Min Yang, Shawn M. Burgess, Rainer K. Brachmann, and David Haussler. 2007. “Species-Specific Endogenous Retroviruses Shape the Transcriptional Network of the Human Tumor Suppressor Protein p53.” *Proceedings of the National Academy of Sciences of the United States of America* 104 (47): 18613–18.

Weinstein, John N., The Cancer Genome Atlas Research Network, Eric A. Collisson, Gordon B. Mills, Kenna R. Mills Shaw, Brad A. Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M. Stuart. 2013. “The Cancer Genome Atlas Pan-Cancer Analysis Project.” *Nature Genetics*. https://doi.org/10.1038/ng.2764.

Yeo, Nan Cher, Alejandro Chavez, Alissa Lance-Byrne, Yingleong Chan, David Menn, Denitsa Milanova, Chih-Chung Kuo, et al. 2018. “An Enhanced CRISPR Repressor for Targeted Mammalian Gene Regulation.” *Nature Methods* 15 (8): 611–16.

YiRen, Hu, Yu YingCong, You Sunwu, Li Keqin, Tong Xiaochun, Chen Senrui, Chen Ende, Lin XiZhou, and Chen Yanfan. 2017. “Long Noncoding RNA MALAT1 Regulates Autophagy Associated Chemoresistance via miR-23b-3p Sequestration in Gastric Cancer.” *Molecular Cancer* 16 (1): 174.

You, Jueng Soo, and Peter A. Jones. 2012. “Cancer Genetics and Epigenetics: Two Sides of the Same Coin?” *Cancer Cell* 22 (1): 9–20.

Zheng, Rongbin, Changxin Wan, Shenglin Mei, Qian Qin, Qiu Wu, Hanfei Sun, Chen-Hao Chen, et al. 2019. “Cistrome Data Browser: Expanded Datasets and New Tools for Gene Regulatory Analysis.” *Nucleic Acids Research* 47 (D1): D729–35.

ACKNOWLEDGEMENTS

We thank the University of Colorado Genomics Shared Resource and BioFrontiers Computing core for technical support during this study. A.I. was supported by the National Cancer Center. E.B.C. was supported by the National Institutes of Health (1R35GM128822), the Alfred P. Sloan Foundation, the David and Lucile Packard Foundation, and the Boettcher foundation.

AUTHOR CONTRIBUTIONS

A.I. designed and conducted the bioinformatics analysis. D.M.S. designed and conducted the experiments. A.I. and E.B.C. conceived the study, analyzed the data, and wrote the paper.

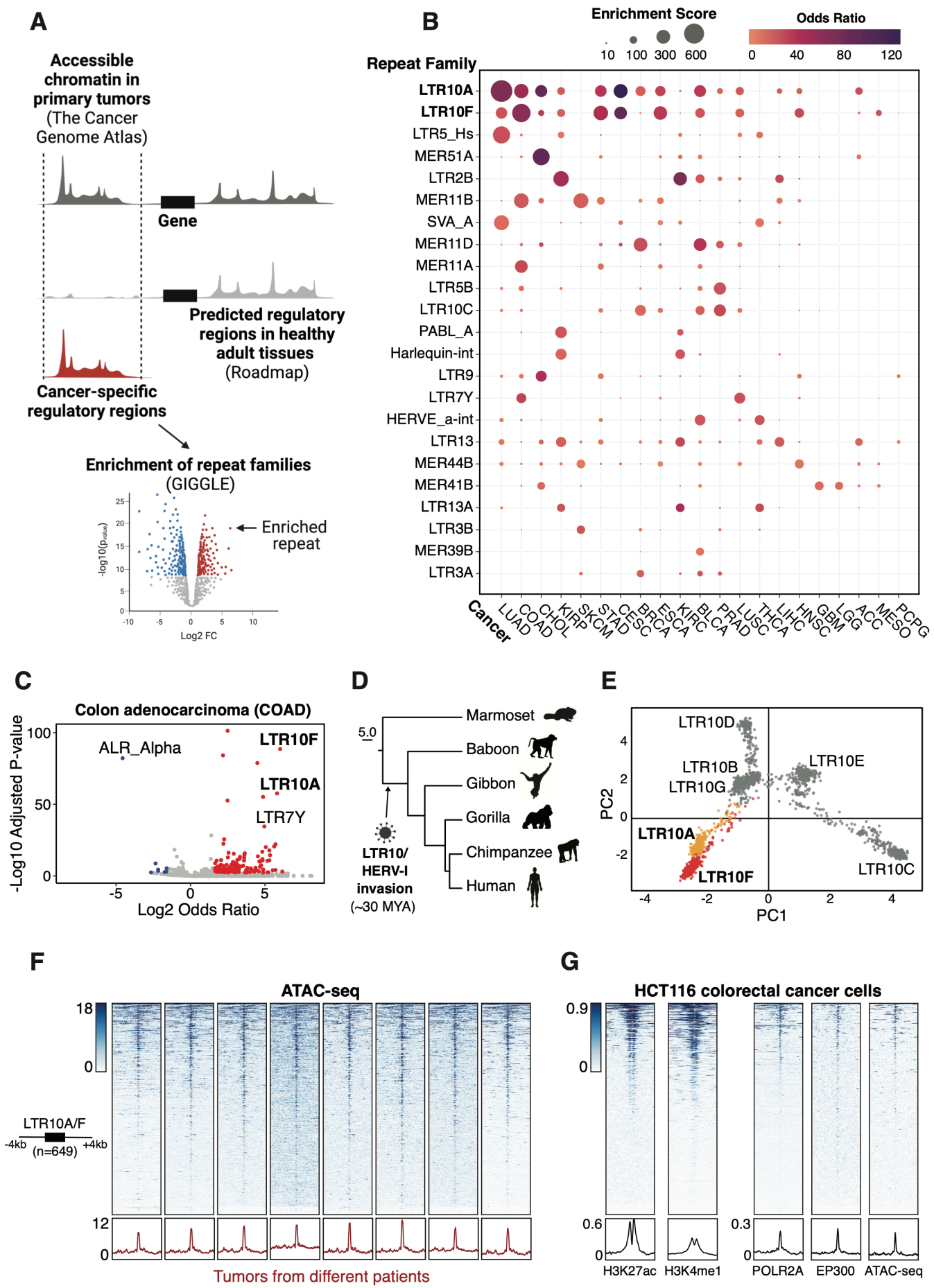
A
Accessible chromatin in primary tumors (The Cancer Genome Atlas)
Gene
Predicted regulatory regions in healthy adult tissues (Roadmap)
Cancer-specific regulatory regions
Enrichment of repeat families (GIGGLE)
Enriched repeat
-log10(p_value)
Log2 FC
B
Enrichment Score
10 100 300 600
Odds Ratio
0 40 80 120
Repeat Family
LTR10A
LTR10F
LTR5_Hs
MER51A
LTR2B
MER11B
SVA_A
MER11D
MER11A
LTR5B
LTR10C
PABL_A
Harlequin-int
LTR9
LTR7Y
HERVE_a-int
LTR13
MER44B
MER41B
LTR13A
LTR3B
MER39B
LTR3A
Cancer
LUAD COAD CHOL KIRP SKCM STAD CESC BRCA ESCA KIRC BLCA PRAD LUSC THCA LIHC HNSC GBM LGG ACC MESO PCPG
C
Colon adenocarcinoma (COAD)
-Log10 Adjusted P-value
ALR_Alpha
LTR10F
LTR10A
LTR7Y
Log2 Odds Ratio
D
Marmoset
Baboon
Gibbon
Gorilla
Chimpanzee
Human
5.0
LTR10/ HERV-I invasion (~30 MYA)
E
LTR10D
LTR10B
LTR10G
LTR10E
LTR10A
LTR10F
LTR10C
PC2
PC1
F
ATAC-seq
LTR10A/F
-4kb
(n=649)
+4kb
Tumors from different patients
G
HCT116 colorectal cancer cells
H3K27ac
H3K4me1
POLR2A
EP300
ATAC-seq

Figure 1: Pan-cancer epigenomic analysis of TE activity. (A) Pipeline to estimate TE family enrichment within cancer-specific regulatory regions. Aggregate ATAC-seq maps associated with each TCGA tumor type were filtered to remove regulatory regions predicted in any normal adult tissues by Roadmap. Cancer-specific accessible chromatin regions were tested for enrichment of 1315 repeat families using GIGGLE. **(B)** Bubble chart summarizing TE family enrichment within cancer-specific ATAC-seq regions across 21 cancer types profiled by TCGA. TE families and cancer types are sorted based on maximum enrichment score. **(C)** Enrichment of TE families within cancer-specific ATAC-seq associated with colon adenocarcinomas (COAD) from TCGA. Significantly enriched TEs are shown in red; depleted TEs are shown in blue. **(D)** Estimated origin of HERV-I elements on the primate phylogeny based on genomic presence or absence. **(E)** Principal component analysis based on sequence similarity of all LTR10 sequences (over 200bp in length) in the human genome. **(F)** Heatmap of representative patient tumor ATAC-seq signals (TCGA patients COAD P053, P012, P002, P025, P004, P016, P001, P049) over the merged set of 649 LTR10A/F elements. Bottom metaprofiles represent average normalized ATAC signal across elements. **(G)** Heatmap of enhancer-associated chromatin marks from HCT116 cells over the merged set of 649 LTR10A/F elements. From left to right: H3K27ac ChIP-seq (GSE97527), H3K4me1 ChIP-seq (GSE101646), POLR2A ChIP-seq (GSE32465), EP300 ChIP-seq (GSE51176), and HCT116 ATAC-seq (GSE126215). Bottom metaprofiles represent the normalized signal across elements.

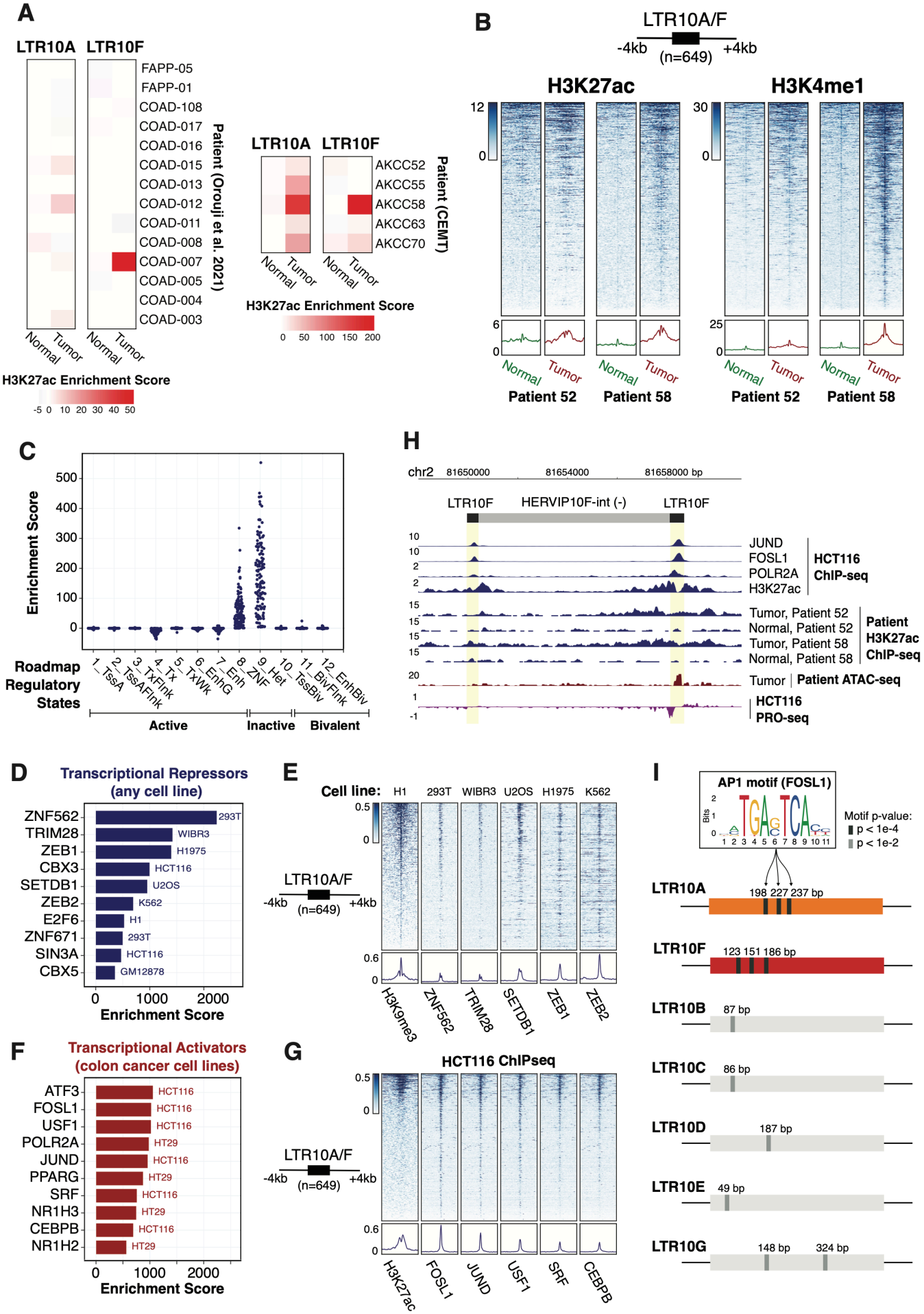
A
LTR10A
LTR10F
FAPP-05
FAPP-01
COAD-108
COAD-017
COAD-016
COAD-015
COAD-013
COAD-012
COAD-011
COAD-008
COAD-007
COAD-005
COAD-004
COAD-003
Patient (Orouji et al. 2021)
Normal
Tumor
H3K27ac Enrichment Score
-5 0 10 20 30 40 50
AKCC52
AKCC55
AKCC58
AKCC63
AKCC70
Patient (CEMT)
0 50 100 150 200
B
LTR10A/F
-4kb
(n=649)
+4kb
H3K27ac
H3K4me1
Patient 52
Patient 58
C
Enrichment Score
Roadmap Regulatory States
1_TssA
2_TssAFlnk
3_TxFlnk
4_Tx
5_TxWk
6_EnhG
7_Enh
8_ZNF
9_Het
10_TssBiv
11_BivFlnk
12_EnhBiv
Active
Inactive
Bivalent
H
chr2
81650000
81654000
81658000 bp
LTR10F
HERVIP10F-int (-)
JUND
FOSL1
POLR2A
H3K27ac
HCT116 ChIP-seq
Tumor, Patient 52
Normal, Patient 52
Tumor, Patient 58
Normal, Patient 58
Patient H3K27ac ChIP-seq
Tumor
Patient ATAC-seq
HCT116 PRO-seq
D
Transcriptional Repressors (any cell line)
ZNF562
TRIM28
ZEB1
CBX3
SETDB1
ZEB2
E2F6
ZNF671
SIN3A
CBX5
293T
WIBR3
H1975
HCT116
U2OS
K562
H1
293T
HCT116
GM12878
Enrichment Score
E
Cell line:
H1
293T
WIBR3
U2OS
H1975
K562
LTR10A/F
-4kb
(n=649)
+4kb
H3K9me3
ZNF562
TRIM28
SETDB1
ZEB1
ZEB2
F
Transcriptional Activators (colon cancer cell lines)
ATF3
FOSL1
USF1
POLR2A
JUND
PPARG
SRF
NR1H3
CEBPB
NR1H2
HCT116
HT29
Enrichment Score
G
HCT116 ChIPseq
LTR10A/F
-4kb
(n=649)
+4kb
H3K27ac
FOSL1
JUND
USF1
SRF
CEBPB
I
AP1 motif (FOSL1)
Bits
Motif p-value:
p < 1e-4
p < 1e-2
LTR10A
198 227 237 bp
LTR10F
123 151 186 bp
LTR10B
87 bp
LTR10C
86 bp
LTR10D
187 bp
LTR10E
49 bp
LTR10G
148 bp
324 bp

Figure 2: Regulatory activity of LTR10 in tumor and normal cells. **(A)** Enrichment of LTR10A/F elements in H3K27ac-defined regulatory regions in patient-matched tumor and normal samples. ChIP-Seq data obtained from Orouhi et al, 2021, and the CEMT Canadian Epigenome Project. **(B)** Heatmap of H3K27ac and H3K4me1 ChIP-seq signals from tumor and normal samples from colorectal cancer patients (CEMT patients AKCC52 and AKCC58) over LTR10A/F elements. Bottom metaprofiles represent average normalized ChIP signal. **(C)** Dotplot showing enrichment scores of LTR10A/F elements for all Roadmap tissues (n=127), across different regulatory states defined by Roadmap. **(D)** Transcriptional repressors associated with LTR10A/F elements, ranked by enrichment score. **(E)** Heatmap of ChIP-seq signal from H3K9me3 and repressive factors, over LTR10A/F elements. From left to right: H3K9me3 ChIP-seq (GSE16256), ZNF562 ChIP-seq (GSE78099), TRIM28 ChIP-seq (GSE84382), SETDB1 ChIP-seq (GSE31477), ZEB1 (GSE106896), and ZEB2 ChIP-seq (GSE91406). **(F)** Transcriptional activators associated with LTR10A/F elements, ranked by enrichment score. **(G)** Heatmap of ChIP-seq signal from H3K27ac and activating transcription factors in HCT116 cells, over LTR10A/F elements. From left to right: H3K27ac ChIP-seq (GSE96299), as well as ChIP-seq for FOSL1, JUND, USF1, SRF and CEBPB (all from GSE32465). **(H)** Genome browser view of a full-length HERV-I element. From top to bottom: JUND and FOSL1 ChIP-seq (GSE32465), POL2RA ChIP-seq (GSE47938), H3K27ac ChIP-seq (GSE73319), H3K27ac ChIP-seq from matched tumor/normal samples from the CEMT Canadian Epigenome Project (patients AKCC52 and AKCC58), tumor ATAC-seq from TCGA (patient COAD P022), and HCT116 PRO-seq (GSE129501). **(I)** Schematic of AP1 motif locations for LTR10 consensus sequences from each subfamily. Sequence logo for AP1 motif FOSL1 (MA0477.1 from JASPAR) is shown, and predicted motif locations are marked on each consensus.

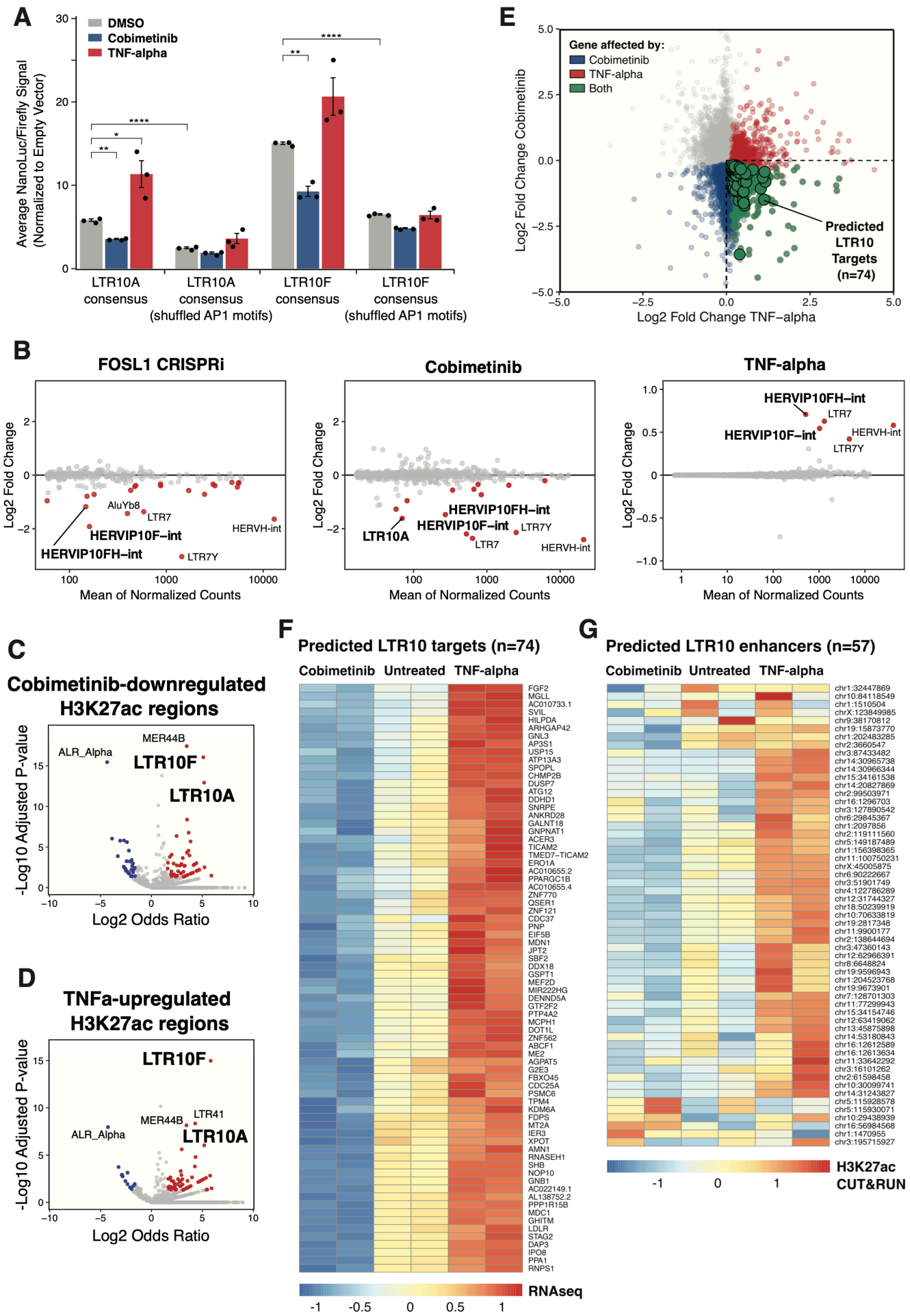
A
DMSO
Cobimetinib
TNF-alpha
Average NanoLuc/Firefly Signal (Normalized to Empty Vector)
LTR10A consensus
LTR10A consensus (shuffled AP1 motifs)
LTR10F consensus
LTR10F consensus (shuffled AP1 motifs)
E
Gene affected by:
Cobimetinib
TNF-alpha
Both
Log2 Fold Change Cobimetinib
Log2 Fold Change TNF−alpha
Predicted LTR10 Targets (n=74)
B
FOSL1 CRISPRi
Cobimetinib
TNF-alpha
Log2 Fold Change
Mean of Normalized Counts
AluYb8
LTR7
HERVIP10F−int
HERVIP10FH−int
LTR7Y
HERVH-int
LTR10A
C
Cobimetinib-downregulated H3K27ac regions
-Log10 Adjusted P-value
Log2 Odds Ratio
ALR_Alpha
MER44B
LTR10F
LTR10A
D
TNFa-upregulated H3K27ac regions
LTR41
F
Predicted LTR10 targets (n=74)
Cobimetinib Untreated TNF-alpha
RNAseq
G
Predicted LTR10 enhancers (n=57)
H3K27ac CUT&RUN

Figure 3: Control of LTR10 activity by AP1/MAPK signaling. (A) Luciferase reporter assays of LTR10A/F consensus sequences, including sequence variants containing shuffled AP1 motifs. Reporter activity was measured in HCT116 cells treated with DMSO, cobimetinib, or TNF-alpha for 24 hrs. Significance was assessed by 2-tailed student's t-test. **(B)** MA plots of TE families showing significant differential expression in HCT116 cells subject to FOSL1 silencing, 24hr cobimetinib treatment, or 24hr TNF-alpha treatment, based on RNA-seq. Dots are colored in red if they are significant (adjusted $p<0.05$, $log2FC<0$ for FOSL1/cobimetinib and $log2FC>0$ for TNFa). **(C)** Volcano plot showing TE family enrichment in the set of H3K27ac regions significantly downregulated by cobimetinib. **(D)** Volcano plot showing TE family enrichment in the set of H3K27ac regions significantly upregulated by TNF-alpha. **(E)** Scatterplot showing MAPK-regulated gene expression changes in response to either cobimetinib or TNF-alpha treatment for 24 hrs. Genes are colored if they are significantly downregulated by cobimetinib (blue), significantly upregulated by TNF-alpha (red), or significantly affected by both treatments in the expected direction (green). Larger bubbles highlight predicted LTR10 target genes (n=74). **(F)** Heatmap of normalized RNA-seq expression values for the 74 predicted LTR10 target genes for each treatment replicate. **(G)** Heatmap of normalized H3K27ac CUT&RUN signal for 57 LTR10 elements predicted to regulate the 74 target genes for each treatment replicate.

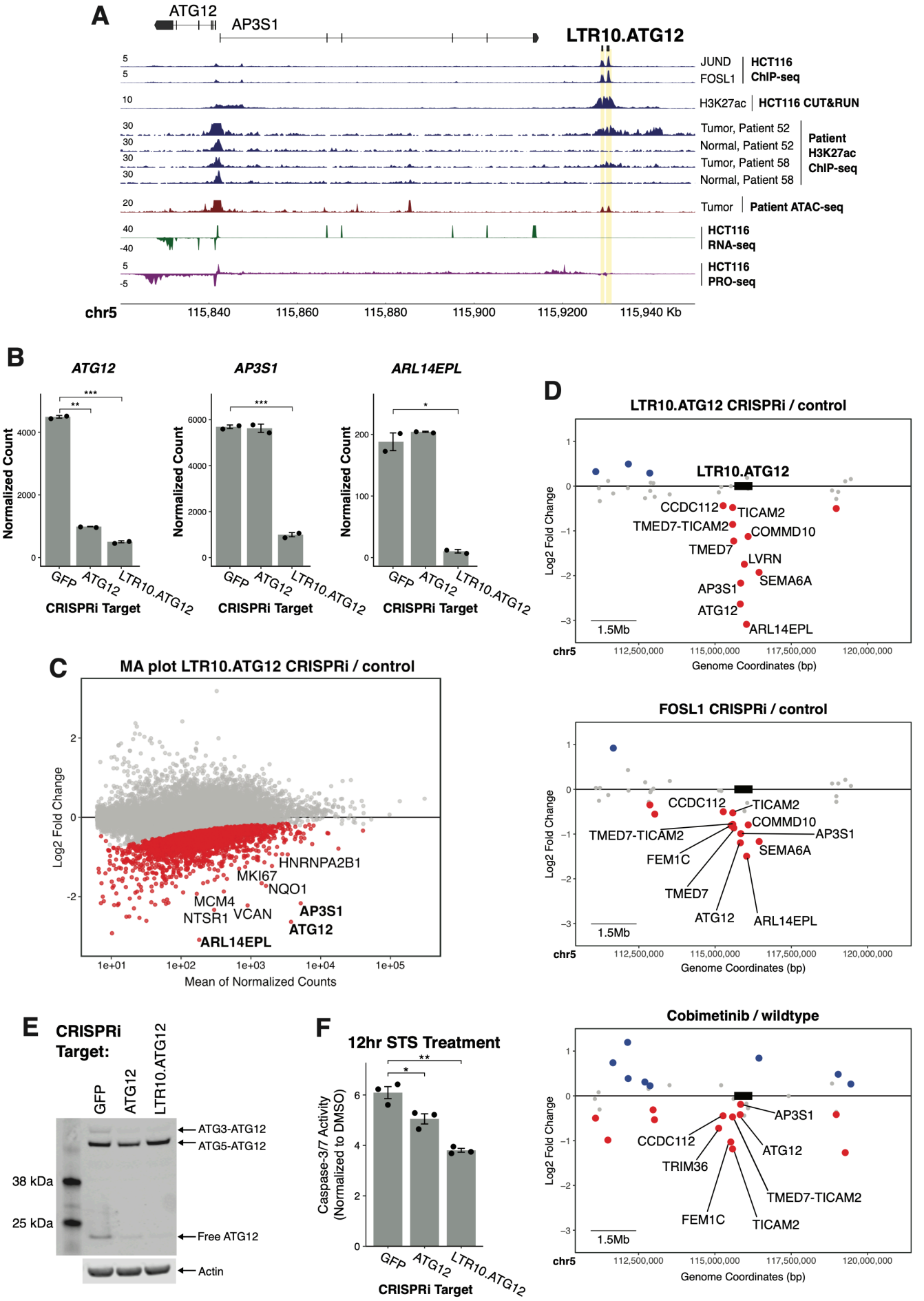
A
ATG12
AP3S1
LTR10.ATG12
JUND
FOSL1
HCT116 ChIP-seq
H3K27ac
HCT116 CUT&RUN
Tumor, Patient 52
Normal, Patient 52
Tumor, Patient 58
Normal, Patient 58
Patient H3K27ac ChIP-seq
Tumor
Patient ATAC-seq
HCT116 RNA-seq
HCT116 PRO-seq
chr5
115,840
115,860
115,880
115,900
115,9200
115,940 Kb
B
ATG12
AP3S1
ARL14EPL
Normalized Count
GFP
ATG12
LTR10.ATG12
CRISPRi Target
C
MA plot LTR10.ATG12 CRISPRi / control
Log2 Fold Change
Mean of Normalized Counts
HNRNPA2B1
MKI67
NQO1
MCM4
NTSR1
VCAN
AP3S1
ATG12
ARL14EPL
D
LTR10.ATG12 CRISPRi / control
LTR10.ATG12
CCDC112
TICAM2
TMED7-TICAM2
COMMD10
TMED7
LVRN
SEMA6A
AP3S1
ATG12
ARL14EPL
1.5Mb
Genome Coordinates (bp)
FOSL1 CRISPRi / control
FEM1C
Cobimetinib / wildtype
TRIM36
E
CRISPRi Target:
GFP
ATG12
LTR10.ATG12
ATG3-ATG12
ATG5-ATG12
38 kDa
25 kDa
Free ATG12
Actin
F
12hr STS Treatment
Caspase-3/7 Activity (Normalized to DMSO)
CRISPRi Target

Figure 4: Functional characterization of LTR10.ATG12 in HCT116 cells. (A) Genome browser screenshot of the *ATG12/AP3S1* locus with the LTR10.ATG12 enhancer labeled. From top to bottom: JUND and FOSL1 ChIP-seq (GSE32465), H3K27ac CUT&RUN (in-house), H3K27ac ChIP-seq from matched tumor/normal samples from the CEMT Canadian Epigenome Project (patients AKCC52 and AKCC58), tumor ATAC-seq from TCGA (patient COAD P022), HCT116 RNA-seq (in-house), and HCT116 PRO-seq (GSE129501). **(B)** Normalized RNA-seq expression values of *ATG12* and *AP3S1* in dCas9-KRAB-MeCP2 HCT116 cells stably transfected with gRNAs targeting the *ATG12* transcription start site, the LTR10.ATG12 element, or non-targeting (GFP) control. **(C)** MA plot showing global gene expression changes in cells in response to silencing LTR10.ATG12. **(D)** Scatterplot of gene expression changes in the locus containing the LTR10.ATG12 element, associated with i) silencing LTR10.ATG12, ii) silencing *FOSL1*, or iii) cobimetinib treatment. Significantly downregulated genes are shown in red; significantly upregulated genes are shown in blue. Significantly downregulated genes located within 1.5 MB of the targeted element are labeled (element box not drawn to scale). **(E)** Immunoblot of endogenous ATG12 in each CRISPRi cell line. Different ATG12 conjugate forms are labeled. **(F)** Caspase-3/7 activity after 12 hrs staurosporine (STS) treatment, measured by the Caspase-Glo 3/7 assay. Treatments were performed in triplicate and signal for each cell line was normalized to signal from DMSO-treatment.

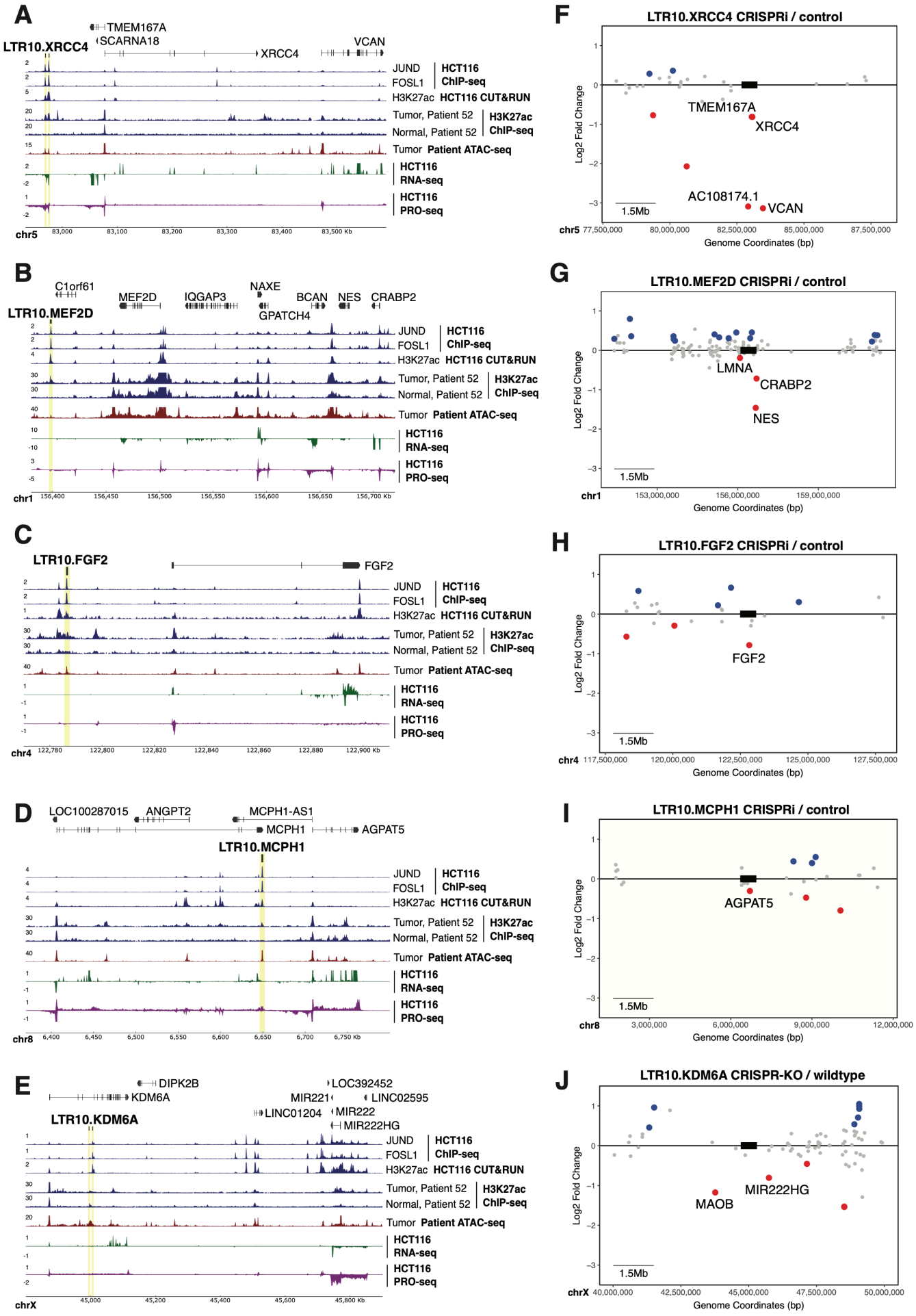
A
LTR10.XRCC4
TMEM167A
SCARNA18
XRCC4
VCAN
JUND
FOSL1
HCT116 ChIP-seq
H3K27ac HCT116 CUT&RUN
Tumor, Patient 52
Normal, Patient 52
H3K27ac ChIP-seq
Tumor Patient ATAC-seq
HCT116 RNA-seq
HCT116 PRO-seq
chr5
B
LTR10.MEF2D
C1orf61
MEF2D
IQGAP3
NAXE
GPATCH4
BCAN
NES
CRABP2
chr1
C
LTR10.FGF2
FGF2
chr4
D
LOC100287015
ANGPT2
MCPH1-AS1
MCPH1
AGPAT5
LTR10.MCPH1
chr8
E
DIPK2B
KDM6A
LOC392452
MIR221
LINC02595
LINC01204
MIR222
MIR222HG
LTR10.KDM6A
chrX
F
LTR10.XRCC4 CRISPRi / control
Log2 Fold Change
TMEM167A
XRCC4
AC108174.1
VCAN
1.5Mb
chr5
Genome Coordinates (bp)
G
LTR10.MEF2D CRISPRi / control
LMNA
CRABP2
NES
chr1
H
LTR10.FGF2 CRISPRi / control
FGF2
chr4
I
LTR10.MCPH1 CRISPRi / control
AGPAT5
chr8
J
LTR10.KDM6A CRISPR-KO / wildtype
MIR222HG
MAOB
chrX

Figure 5: LTR10 elements control gene expression at multiple loci. (A-E) Screenshots of predicted LTR10-derived enhancers selected for validation. Elements are named based on their predicted proximal target gene (XRCC4, MEF2D, FGF2, MCPH1, and KDM6A). **(F-J)** Scatterplots of local gene expression changes in response to CRISPR perturbation of the targeted LTR10 element. Significantly downregulated genes are shown in red; significantly upregulated genes are shown in blue. Significantly downregulated genes within 1.5 MB of the targeted element are labeled (element box not drawn to scale). Black boxes indicate the targeted LTR10 element in each case (not drawn to scale). For all elements except LTR10.KDM6A, CRISPRi transfection replicates targeting the enhancer were compared against cells transfected with a non-targeting gRNA (GFP) control. For LTR10.KDM6A, multiple CRISPR deletion K/O clones were generated and compared against wild-type HCT116 cells.

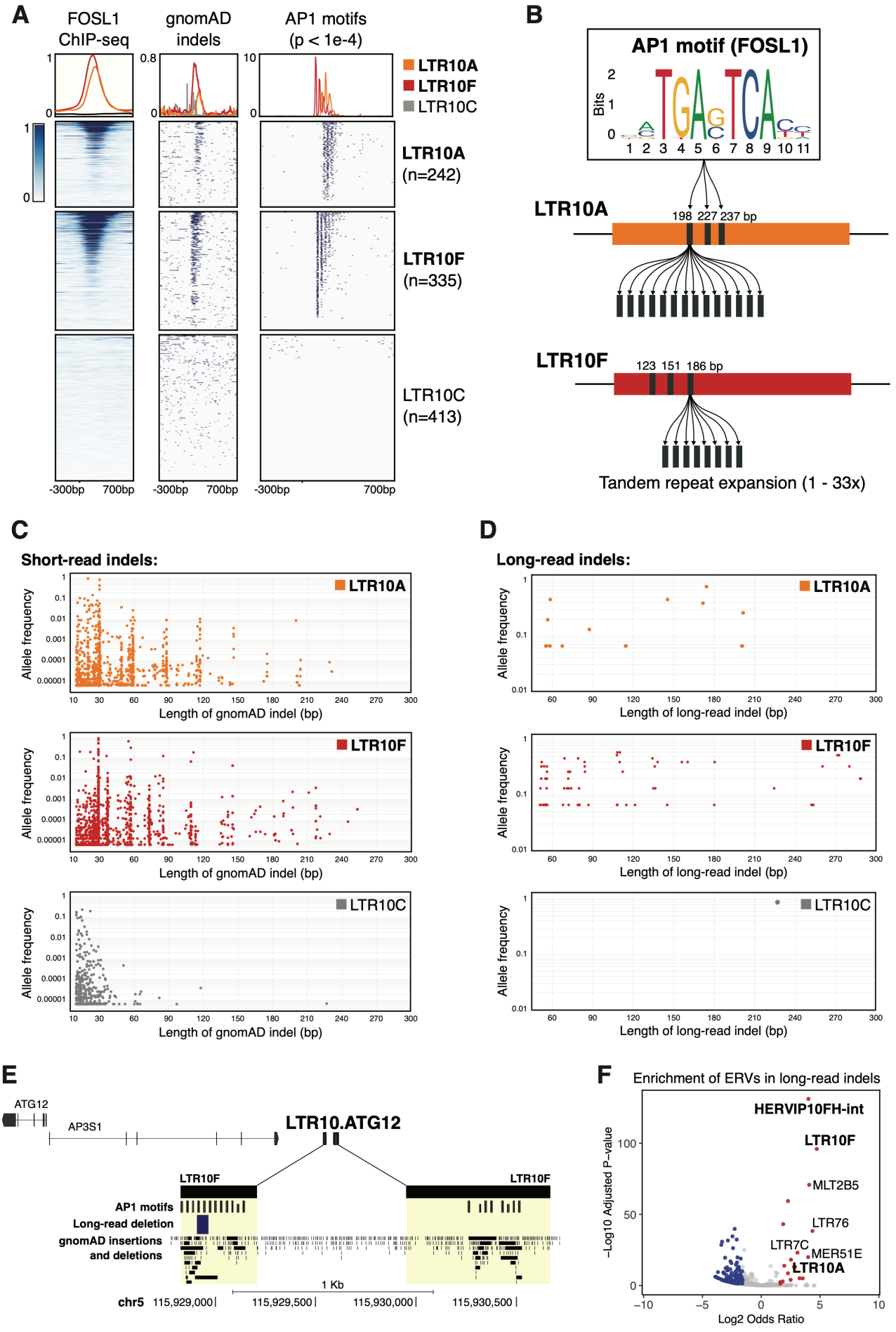
A
FOSL1 ChIP-seq
gnomAD indels
AP1 motifs (p < 1e-4)
LTR10A
LTR10F
LTR10C
LTR10A (n=242)
LTR10F (n=335)
LTR10C (n=413)
-300bp
700bp
B
AP1 motif (FOSL1)
Bits
LTR10A
198 227 237 bp
LTR10F
123 151 186 bp
Tandem repeat expansion (1 - 33x)
C
Short-read indels:
Allele frequency
Length of gnomAD indel (bp)
D
Long-read indels:
Length of long-read indel (bp)
E
ATG12
AP3S1
LTR10.ATG12
AP1 motifs
Long-read deletion
gnomAD insertions and deletions
1 Kb
chr5
115,929,000
115,929,500
115,930,000
115,930,500
F
Enrichment of ERVs in long-read indels
HERVIP10FH-int
LTR10F
MLT2B5
LTR76
LTR7C
MER51E
LTR10A
−Log10 Adjusted P−value
Log2 Odds Ratio

Figure 6: LTR10 repeat instability and polymorphism. (A) Heatmap of FOSL1 ChIP-Seq, gnomAD indels between 10-300bp in length, and AP1 motif matches ($p<1e-4$) across LTR10A, LTR10F, and LTR10C elements. Overlapping elements were removed, retaining 990 LTR10 elements total across the three subfamilies. FOSL1 ChIP-seq was obtained from GSE32465. **(B)** Schematic of variable number tandem repeat (VNTR) region within LTR10A and LTR10F elements. **(C)** Scatterplot of high-confidence gnomAD indels between 10-300 bp in length detected in LTR10A, LTR10F, or LTR10C subfamilies. Each indel is plotted by its length and allele frequency. **(D)** As in (C) but using long-read supported data. **(E)** Genome browser screenshot of LTR10.ATG12 showing AP1 motifs, long-read indels (58bp deletion reported by Quan et al., 2021), and gnomAD indels. **(F)** GIGGLE enrichment of ERVs within long-read indels. Significantly enriched ERVs are shown in red; depleted ERVs are shown in blue.

LIST OF SUPPLEMENTARY MATERIAL

Supplementary Figure S1: Additional figures for the pan-cancer epigenomic analysis of TE activity.

Supplementary Figure S2: Additional figures for the regulatory activity of LTR10 regulatory activity in tumor and normal cells.

Supplementary Figure S3: Additional figures for the control of LTR10 regulatory activity by AP1/MAPK signaling.

Supplementary Figure S4: Additional figures for the CRISPRi silencing of LTR10.ATG12 and ATG12 (TSS).

Supplementary Figure S5: PCR validation of LTR10.KDM6A CRISPR-KO clones.

Supplementary Figure S6: Additional figures showing LTR10 repeat instability and polymorphism.

Supplementary Table 1: TEs enriched in tumor accessible chromatin from TCGA.

Supplementary Table 2: LTR10 elements enriched in Roadmap regulatory regions.

Supplementary Table 3: LTR10 elements enriched in Cistrome Transcription Factors.

Supplementary Table 4: DEseq2 results for cells treated with 24 hr cobimetinib, 24 hr TNF-alpha, or untreated.

Supplementary Table 5: All predicted LTR10 enhancer regions based on the activity by contact model.

Supplementary Table 6: 57 LTR10 enhancers predicted to regulate AP1/MAPK genes.

Supplementary Table 7: DEseq2 results for CRISPRi silencing of LTR10.ATG12 element.

Supplementary Table 8: Primers, gRNAs and LTR10 sequences used for CRISPR experiments and functional validation.

SUPP TABLE LEGENDS

Supplementary Table 1: TEs enriched in tumor accessible chromatin from TCGA. The top 23 TE families (based on GIGGLE enrichment score) are shown.

Supplementary Table 2: LTR10 elements enriched in Roadmap regulatory regions, based on GIGGLE enrichment score. The merged set of 649 LTR10A/F elements was searched against all Roadmap tissues (n=127).

Supplementary Table 3: LTR10 elements enriched in Cistrome Transcription Factors, based on GIGGLE enrichment score. The merged set of 649 LTR10A/F elements was searched against all Cistrome human transcription factors (n=11348).

Supplementary Table 4: 620 MAPK-regulated genes based on DEseq2 results for HCT116 RNAseq with untreated, 24hr cobimetinib-treated and 24hr TNFa-treated cells.

Supplementary Table 5: All predicted LTR10 enhancer regions based on the activity by contact model (*41*), which assigns enhancer-gene targets based on H3K27ac signal and chromatin interaction data.

Supplementary Table 6: 57 LTR10 enhancers predicted to regulate MAPK genes. LTR10 enhancers predicted by the activity by contact model (*41*) were searched for proximity (within 1.5Mb) to genes significantly affected by both cobimetinib and TNF-alpha treatment.

Supplementary Table 7: DEseq2 results comparing the LTR10.ATG12 element silencing to control (GFP). Genes are sorted from lowest to highest adjusted p-value.

Supplementary Table 8: Primers, gRNAs and LTR10 sequences used for CRISPR experiments, functional validation and luciferase assays.

Figure S1A

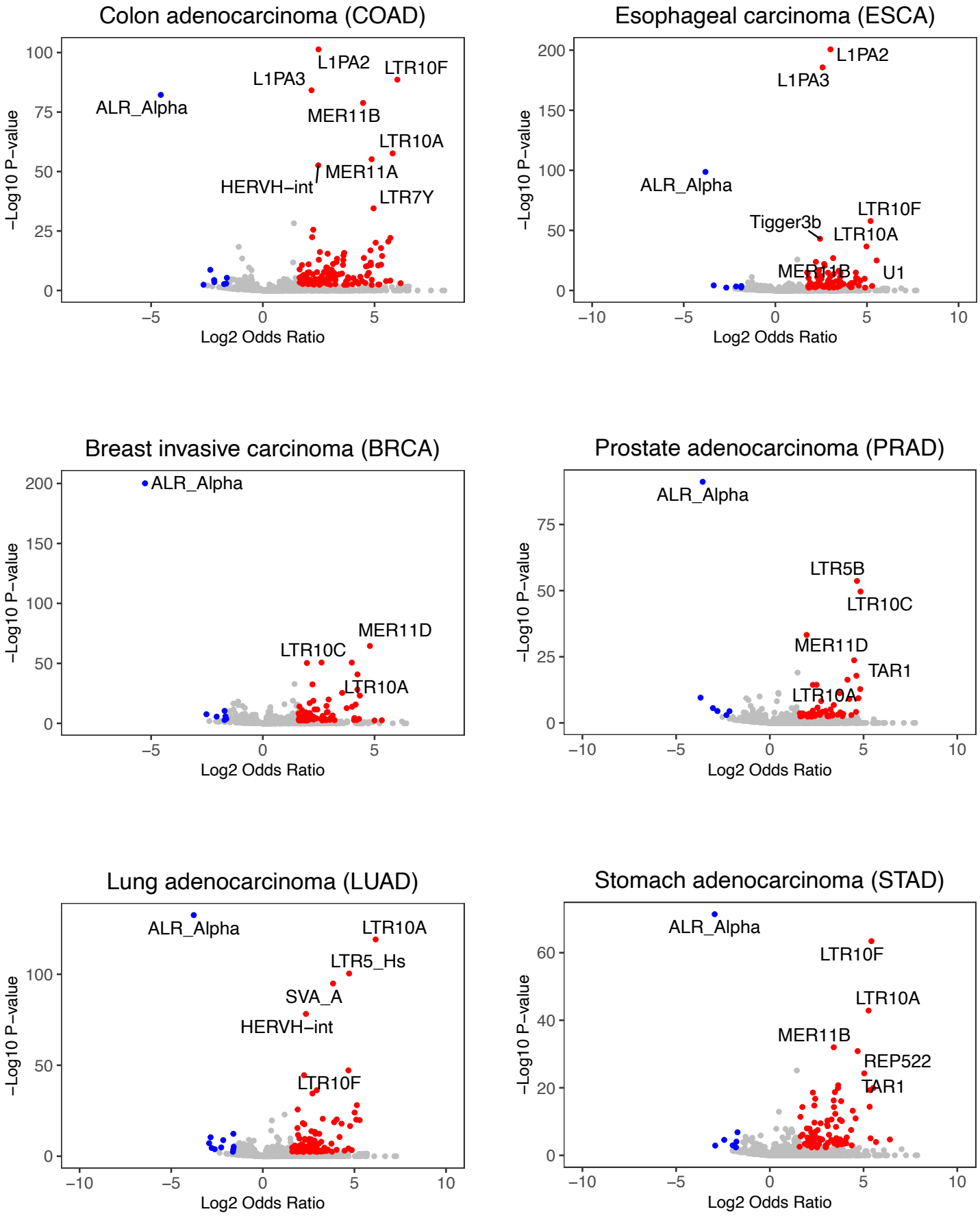

Figure S1A: Enrichment of TE families within cancer-specific ATAC-seq associated with different cancer subtypes from TCGA. Significantly enriched TEs are shown in red; depleted TEs are shown in blue.

Figure S1B

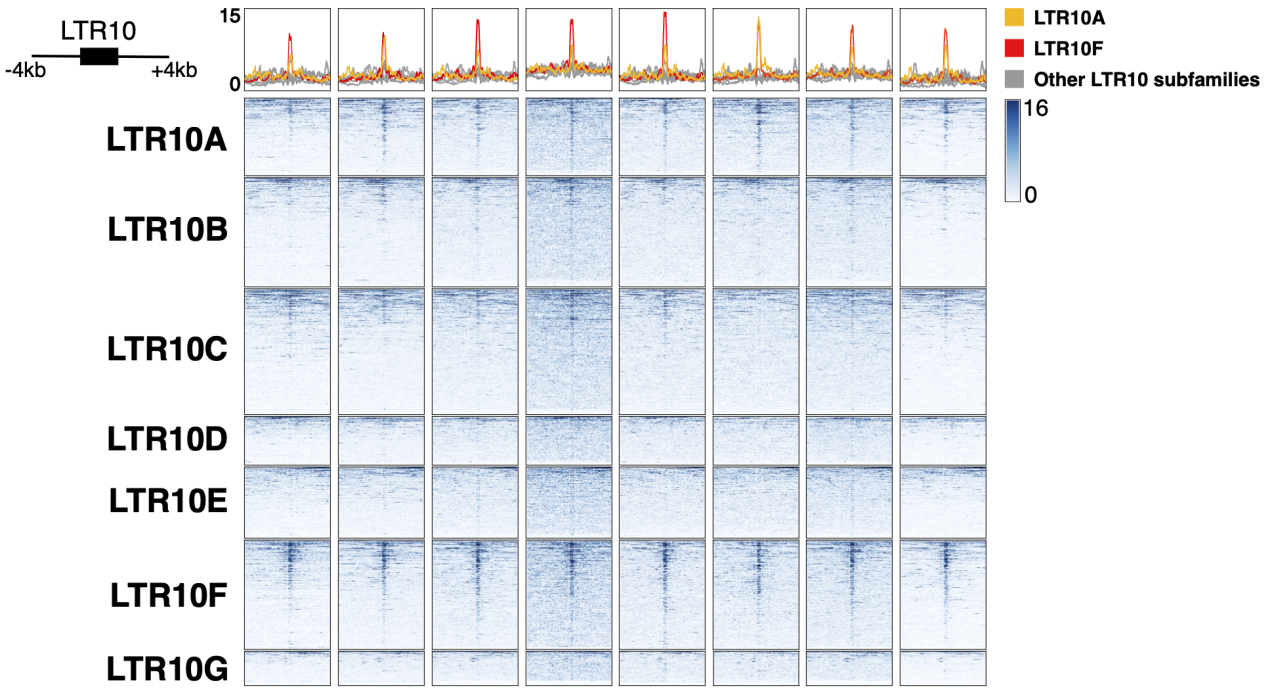

Figure S1B: Heatmap of representative patient tumor ATAC-seq signals (TCGA patients P053, P012, P002, P025, P004, P016, P001, P049) over all LTR10 elements, separated by subfamily. Metaprofiles represent average normalized ATAC signal across elements.

Figure S1C

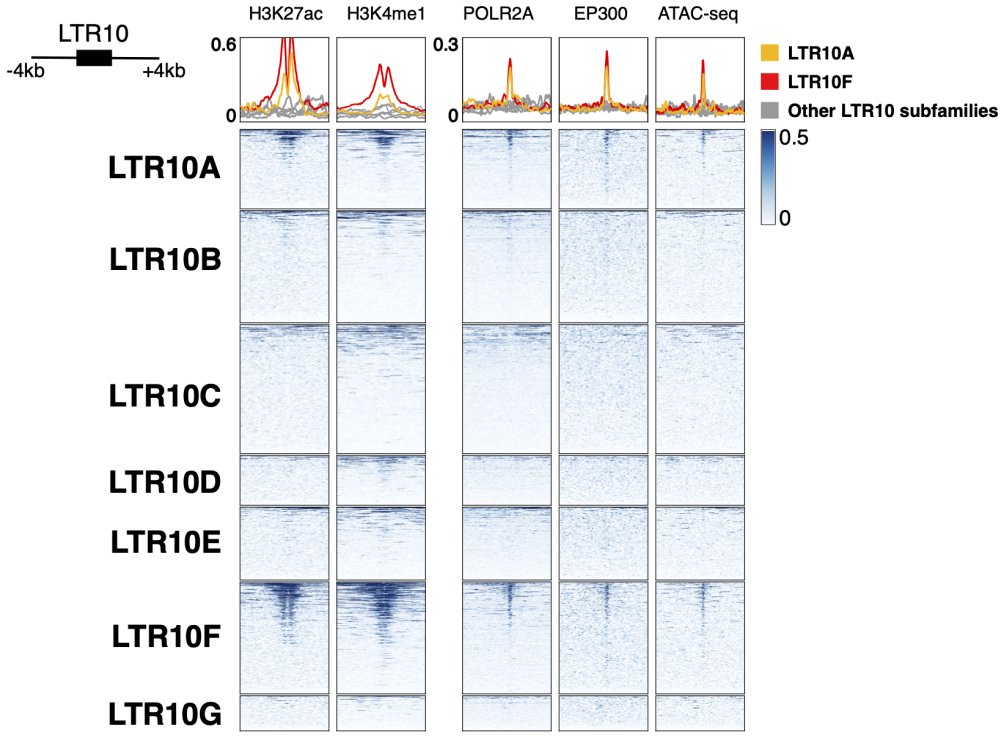

Figure S1C: Heatmap of enhancer-associated chromatin marks from HCT116 cells over all LTR10 elements, separated by subfamily. From left to right: H3K27ac ChIP-seq (GSE97527), H3K4me1 ChIP-seq (GSE101646), POLR2A ChIP-seq (GSE32465), EP300 ChIP-seq (GSE51176), and HCT116 ATAC-seq (GSE126215). Metaprofiles represent the normalized signal across elements.

Figure S1D

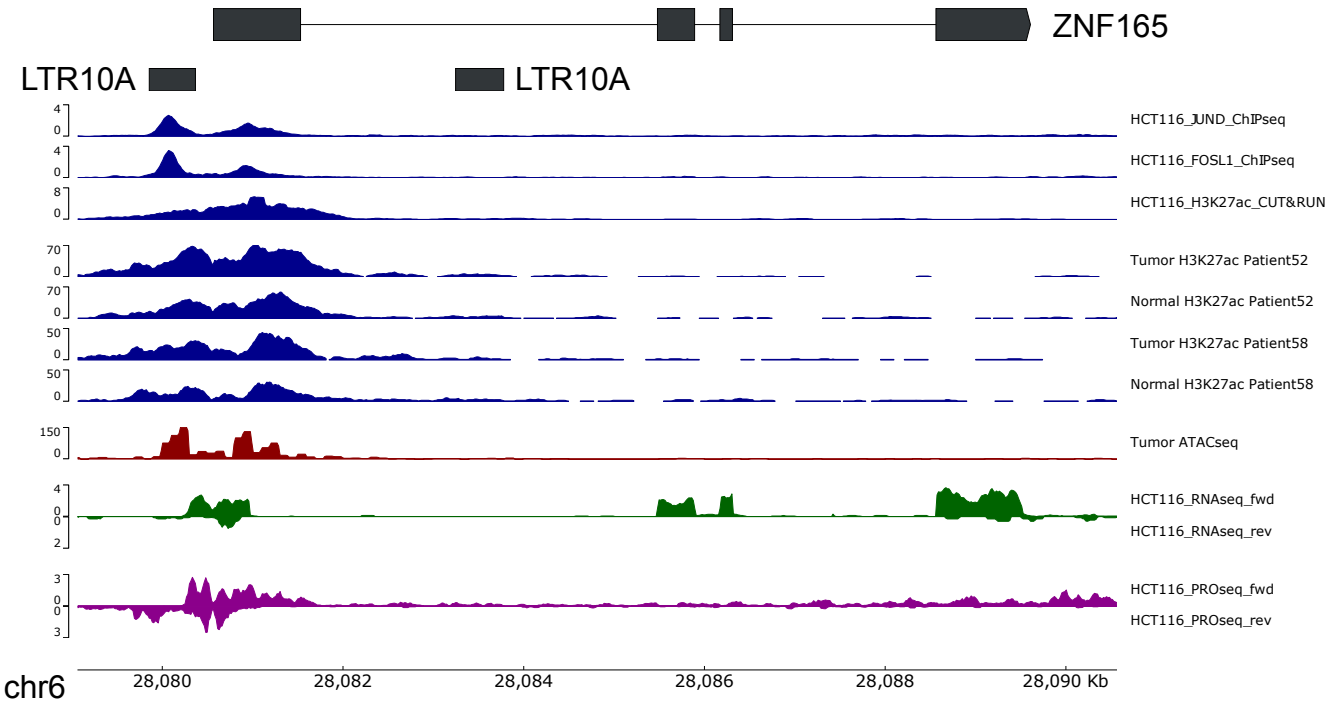

Figure S1D: Genome browser view of an LTR10A element co-opted as a promoter for ZNF165. From top to bottom: JUND and FOSL1 ChIP-seq (GSE32465), H3K27ac CUT&RUN (in-house), H3K27ac ChIP-seq from matched tumor/normal samples from the CEMT Canadian Epigenome Project (patients AKCC52 and AKCC58), tumor ATAC-seq from TCGA (patient COAD P022), HCT116 RNA-seq (in-house), and HCT116 PRO-seq (GSE129501).

Figure S2A

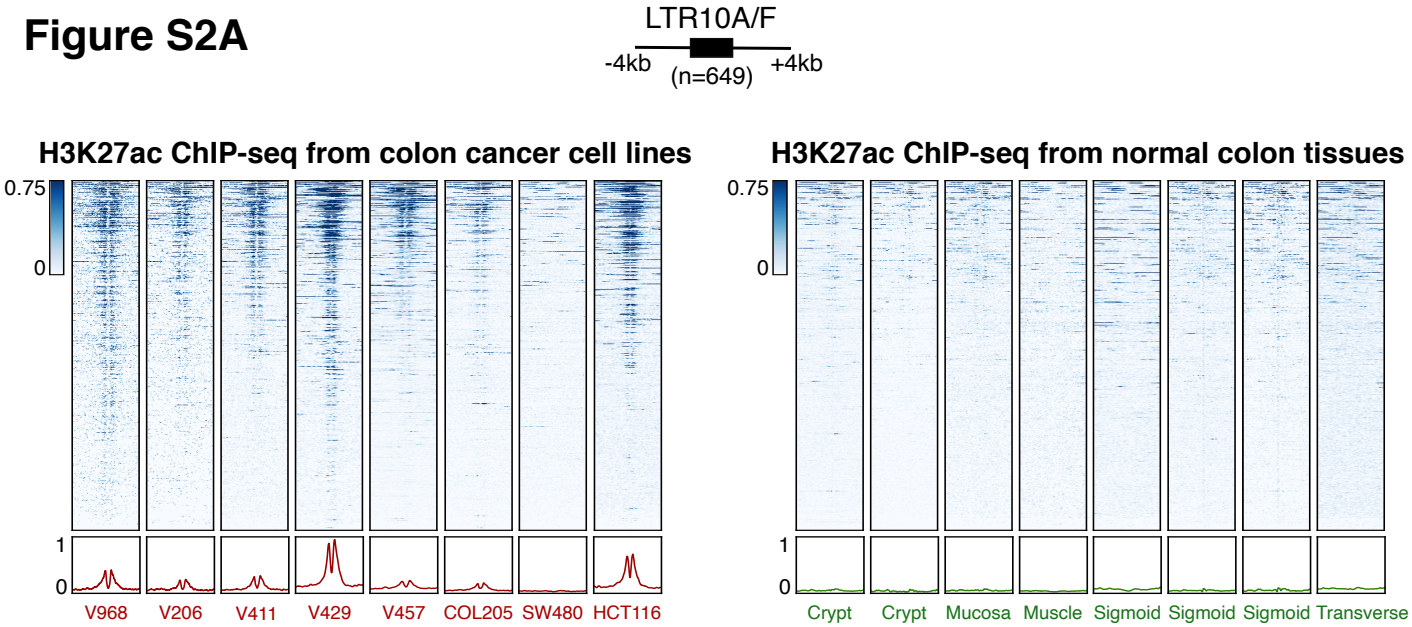

Figure S2A: Heatmap of H3K27ac ChIP-seq from different colorectal cancer cell lines (GSE77737) versus H3K27ac ChIP-seq from normal colon tissues (GSE77737, GSE17312, GSE101136, GSE101031, GSE16256), over the merged set of 649 LTR10A/F elements. Bottom metaprofiles represent the normalized signal across elements.

Figure S2B

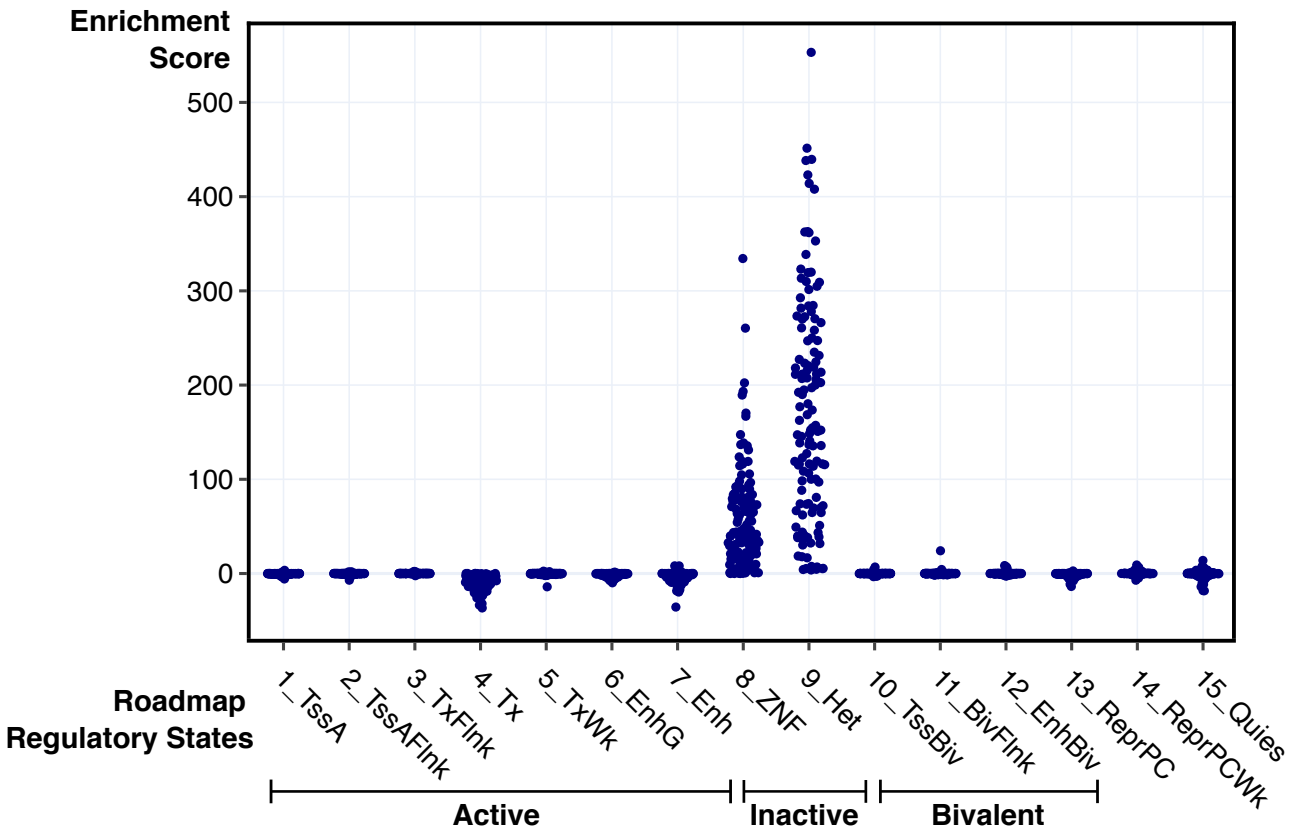

Figure S2B: Dotplot showing enrichment scores of LTR10A/F elements for all Roadmap tissues (N=127), across all fifteen regulatory states defined by Roadmap.

Figure S3A

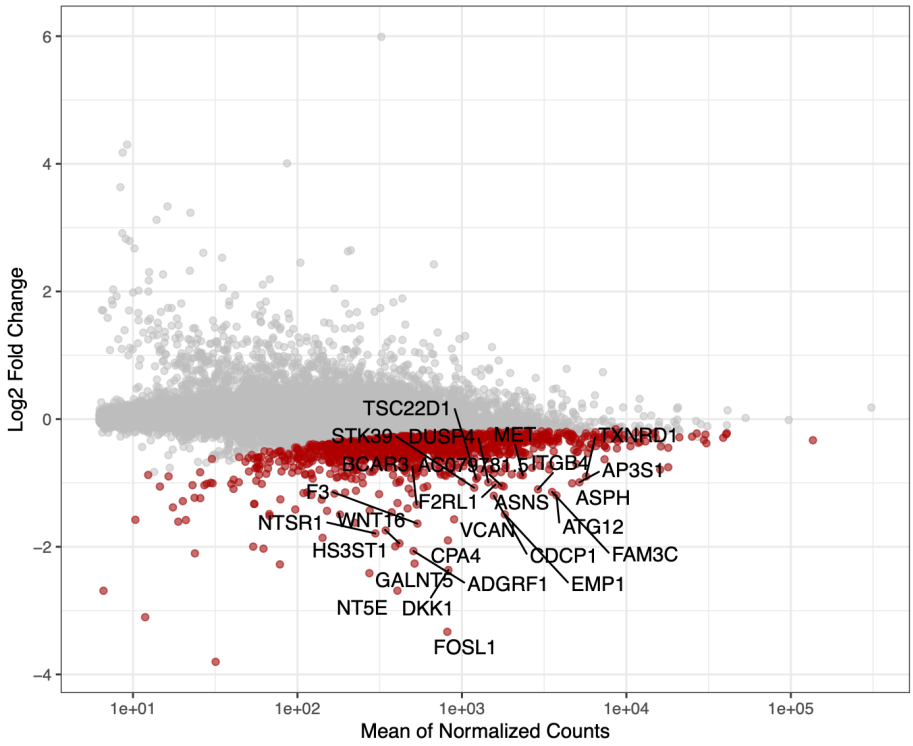

Figure S3A: MA plot showing global gene expression changes in cells in response to silencing FOSL1. Significantly downregulated genes are shown in red.

Figure S3B

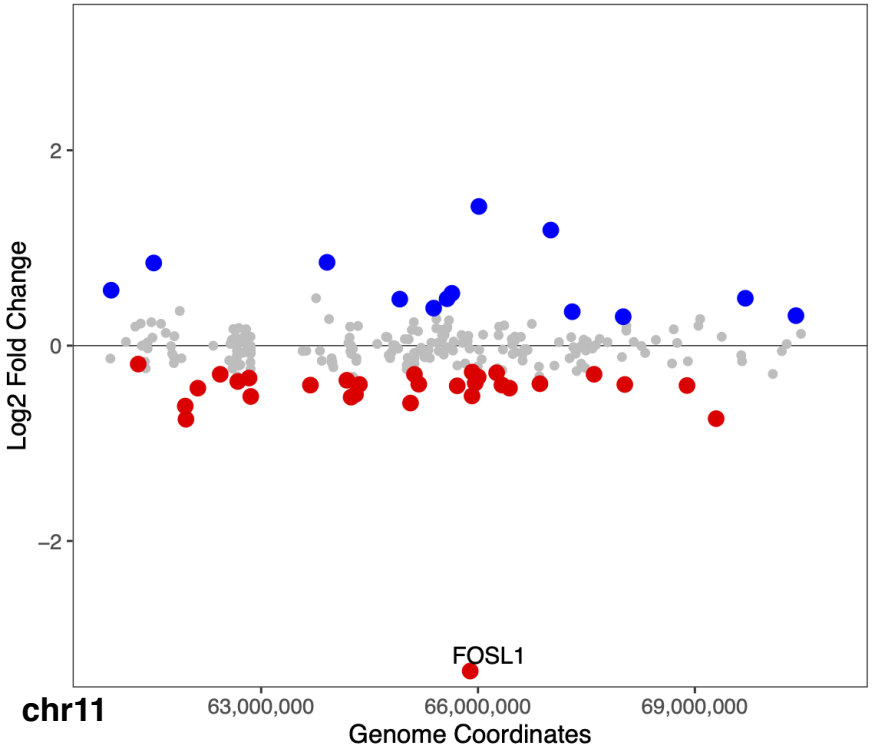

Figure S3B: Scatterplot of gene expression changes in response to silencing FOSL1 in a 10Mb window around the target. Significantly down-regulated genes are shown in red, significantly up-regulated genes are shown in blue. The most significantly downregulated gene (FOSL1) is labelled.

Figure S3C

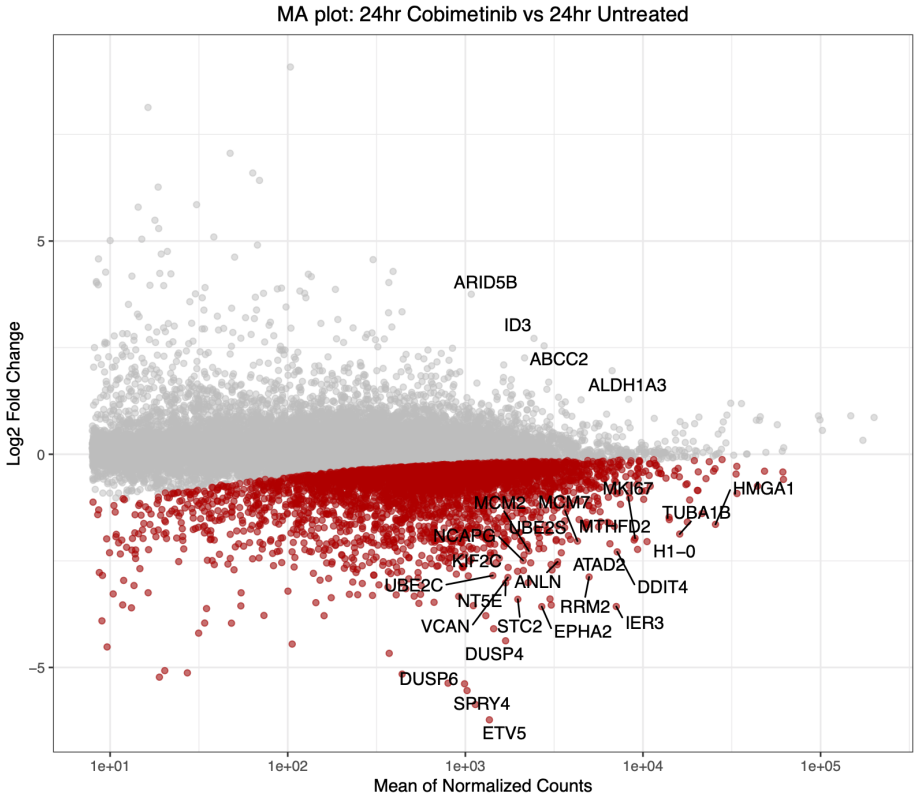

Figure S3C: MA plot showing global gene expression changes in cells in response to 24hr cobimetinib treatment. Significantly downregulated genes are shown in red.

Figure S3D

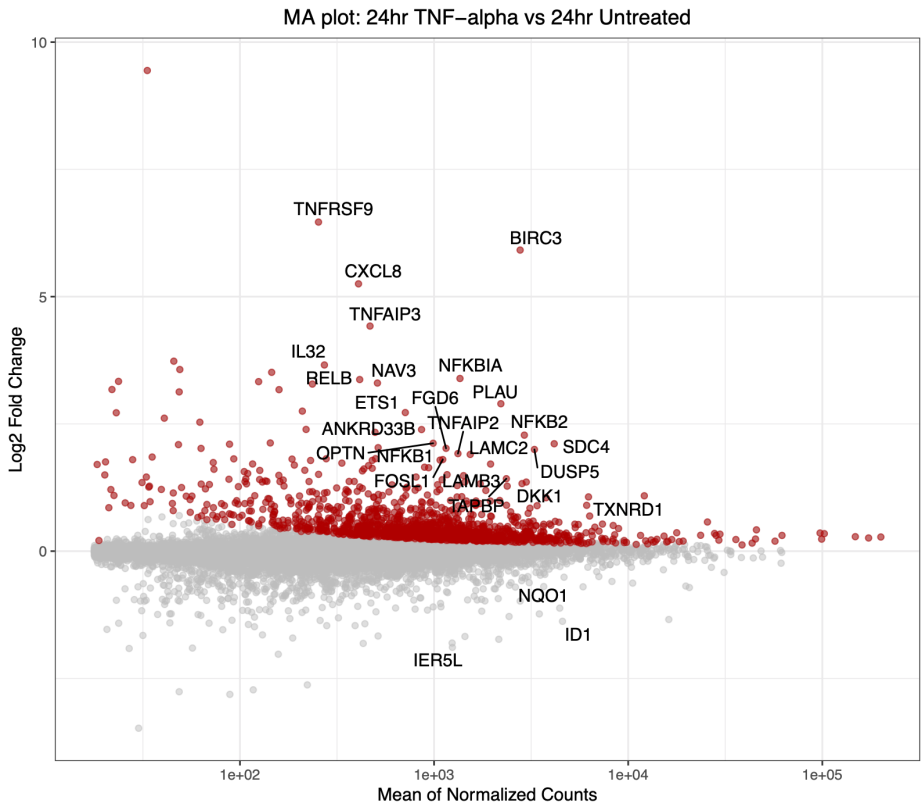

Figure S3D: MA plot showing global gene expression changes in cells in response to 24hr TNF-alpha treatment. Significantly upregulated genes are shown in red.

Figure S3E

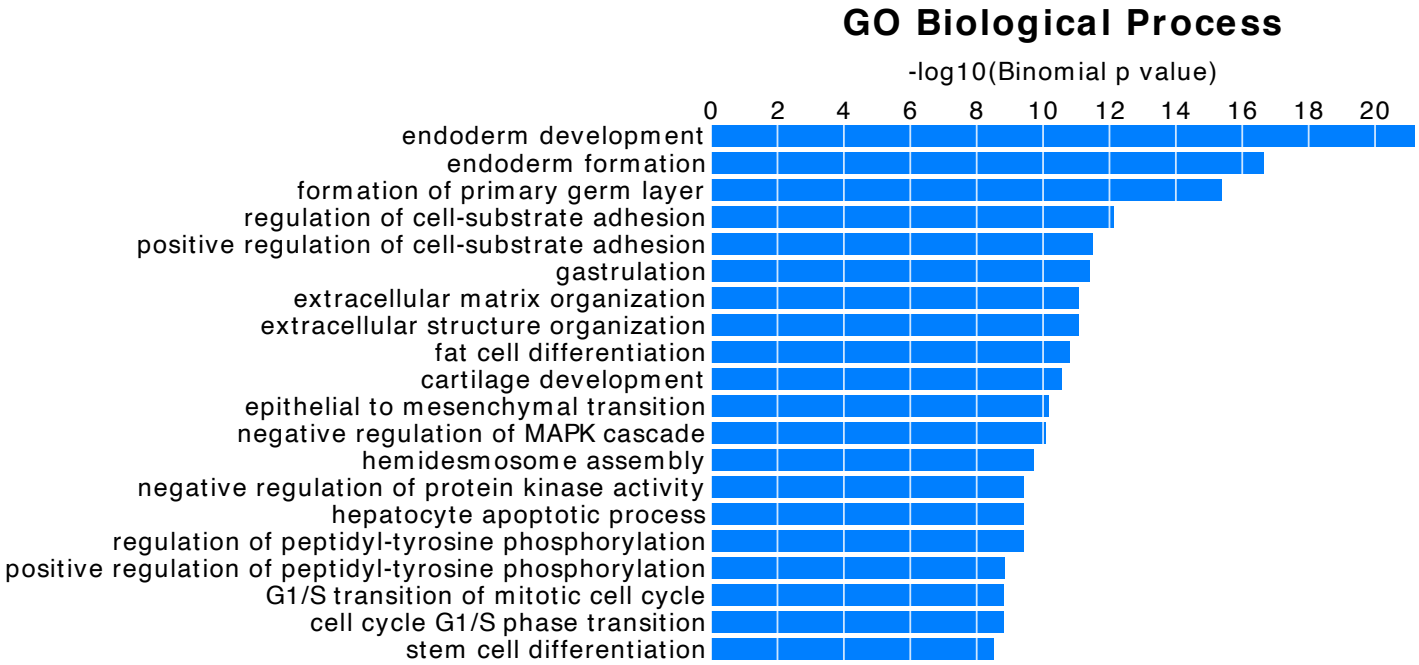

Figure S3E: Bargraph of Gene Ontology (GO) biological processes associated with the set of H3K27ac regions significantly downregulated by cobimetinib (N=1634), as predicted by GREAT.

Figure S3F

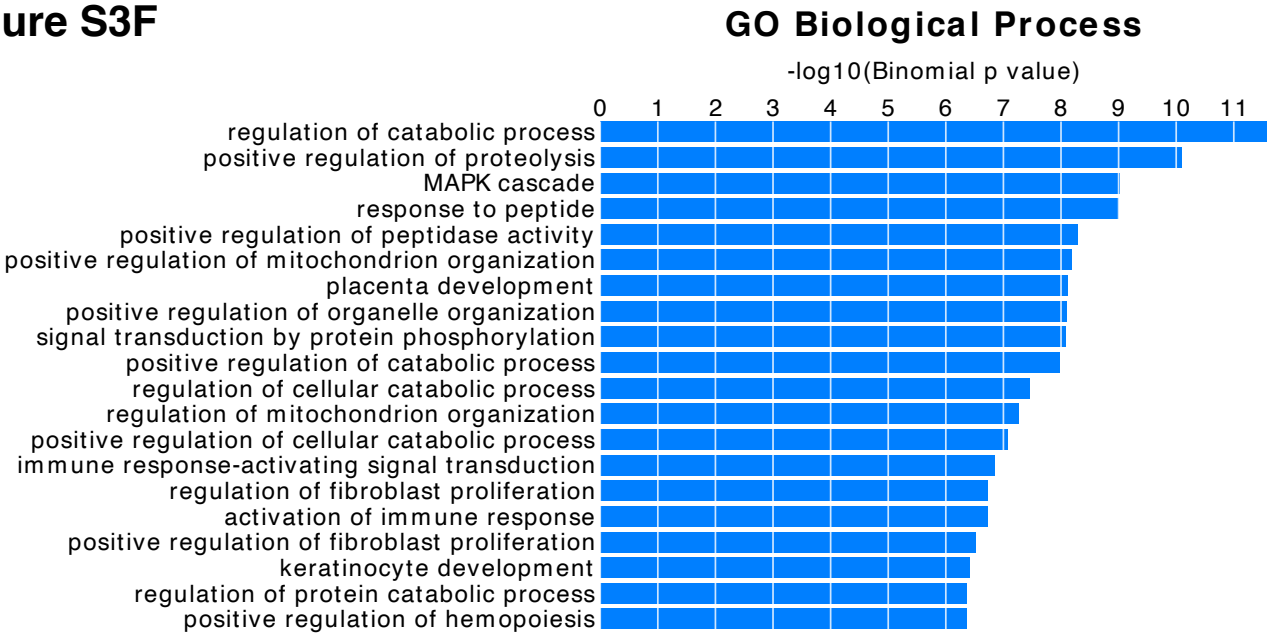

Figure S3F: Bargraph of Gene Ontology (GO) biological processes associated with the set of H3K27ac regions significantly upregulated by TNF-alpha (N=775), as predicted by GREAT.

Figure S3G

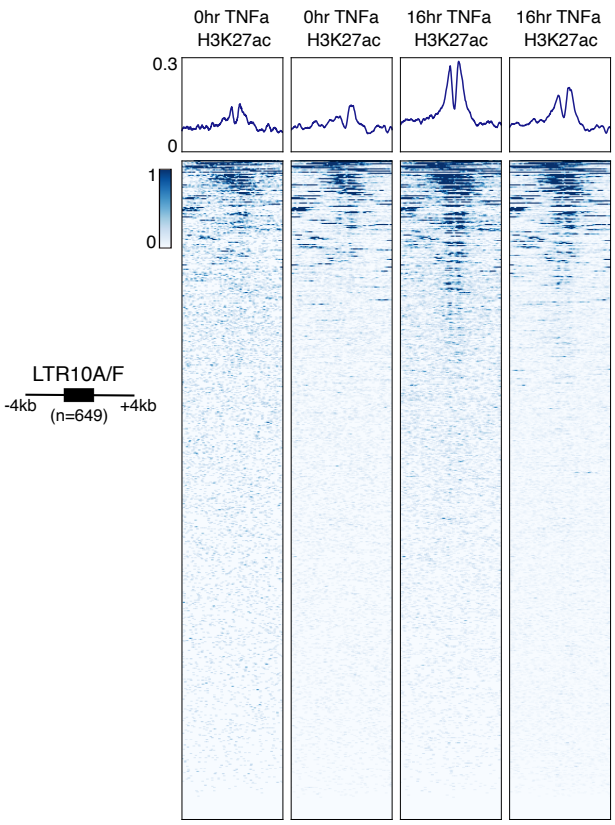

Figure S3G: Heatmap of H3K37ac ChIP-seq from SW480 colon cancer cells, either untreated or treated with TNF-alpha for 16hr, over the merged set of 649 LTR10A/F elements (GSE102796). Metaprofiles represent the normalized signal across elements.

Figure S4A

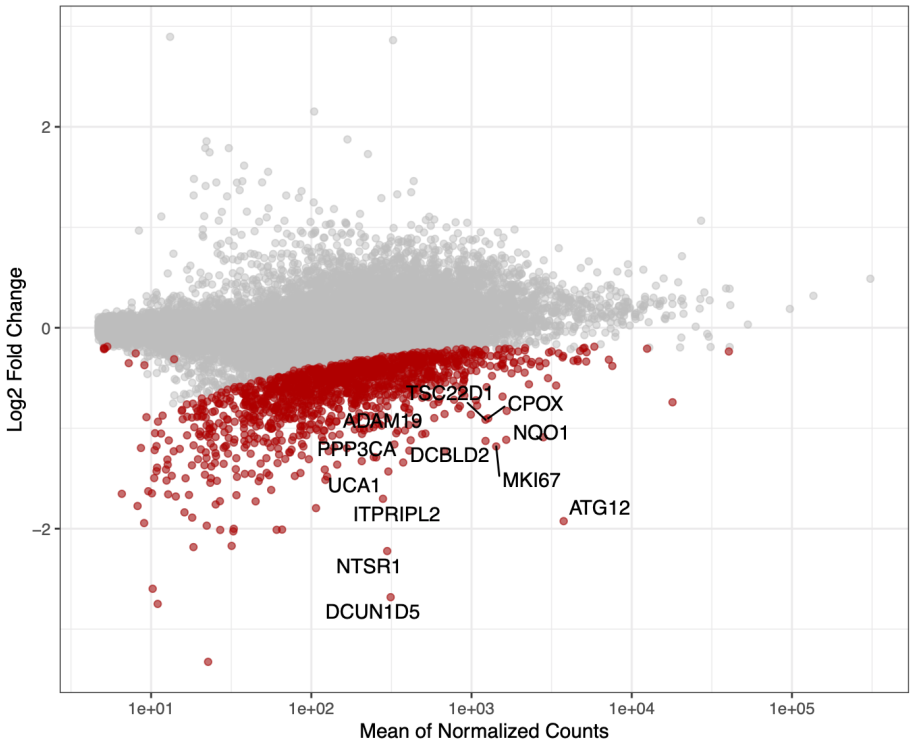

Figure S4A: MA plot showing global gene expression changes in cells in response to silencing ATG12 (TSS). Significantly downregulated genes are shown in red.

Figure S4B

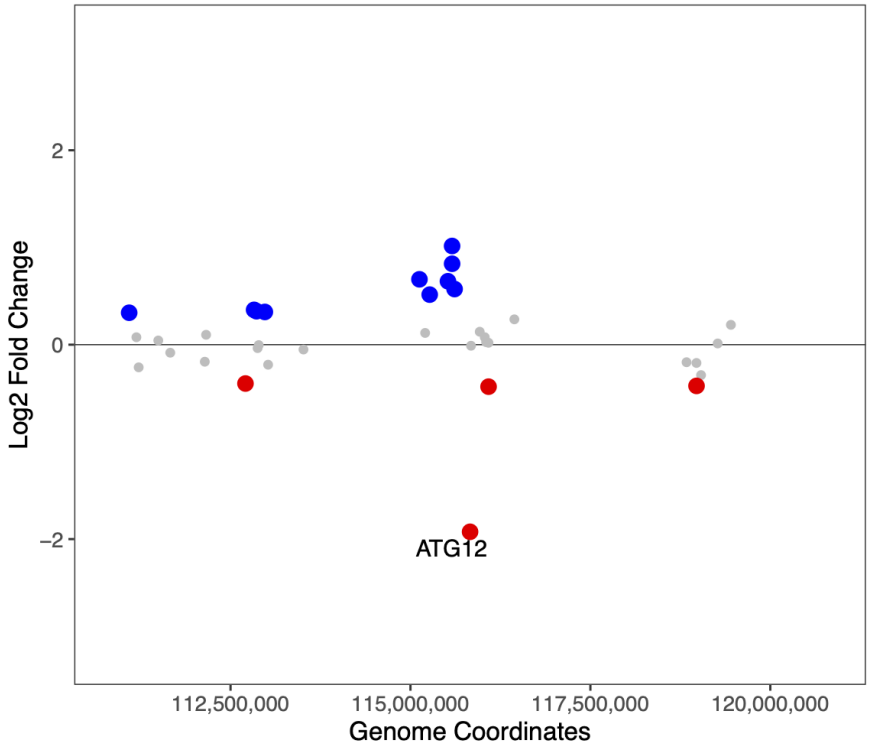

Figure S4B: Scatterplot of gene expression changes in response to silencing ATG12 (TSS) in a 10Mb window around the target. Significantly down-regulated genes are shown in red, significantly up-regulated genes are shown in blue. The most significantly downregulated gene (ATG12) is labelled.

Figure S4C

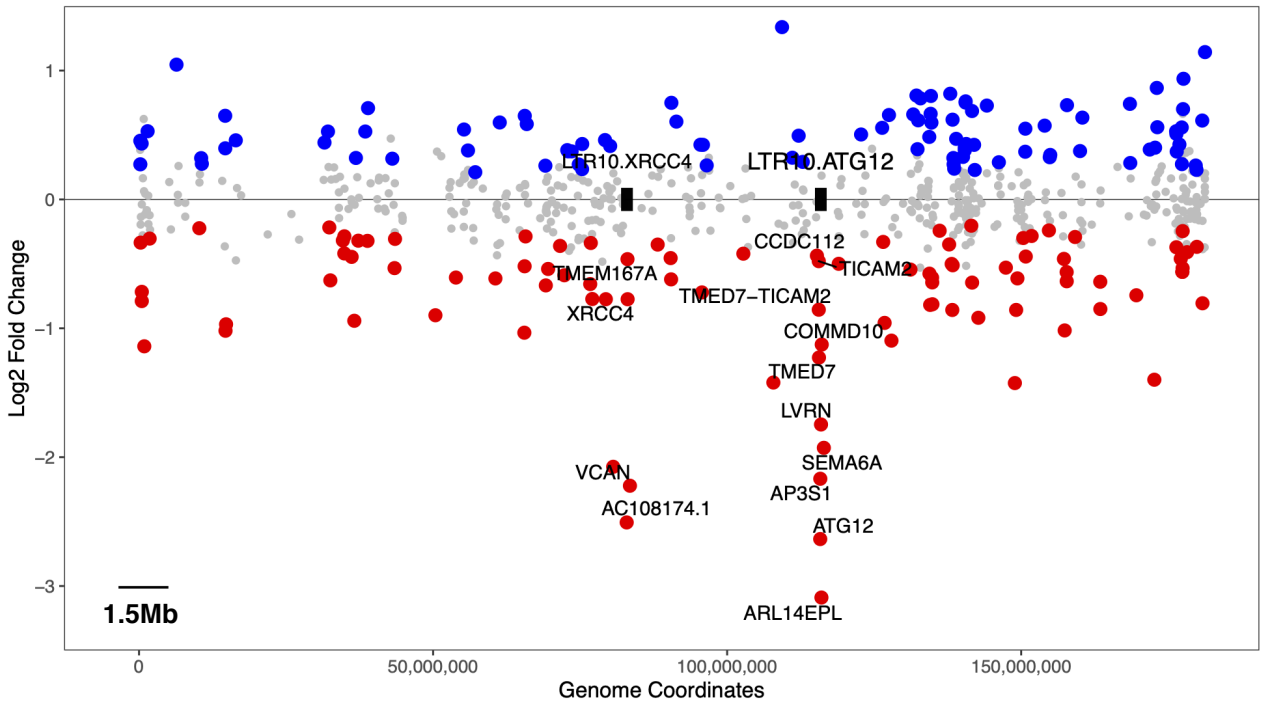

Figure S4C: Scatterplot of gene expression changes in response to silencing LTR10.ATG12, showing all of hg38 chromosome 5. The locations of LTR10.XRCC4 and LTR10.ATG12 are marked (element boxes not drawn to scale). Significantly downregulated genes are shown in red; significantly upregulated genes are shown in blue. Significantly downregulated genes within 1.5 Mb of either LTR10 element are labelled.

Figure S4D

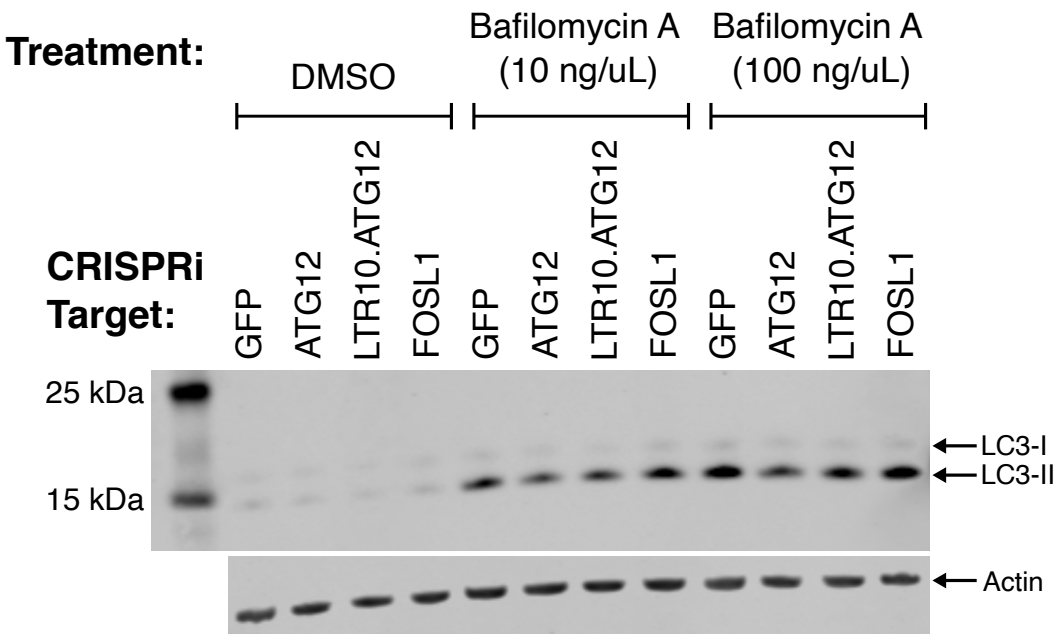

Figure S4D: Immunoblot of LC3-I and LC3-II in each CRISPRi cell line, after treating cells for 6 hr with DMSO, bafilomycin A (10 ng/uL), or bafilomycin A (100 ng/uL).

Figure S5

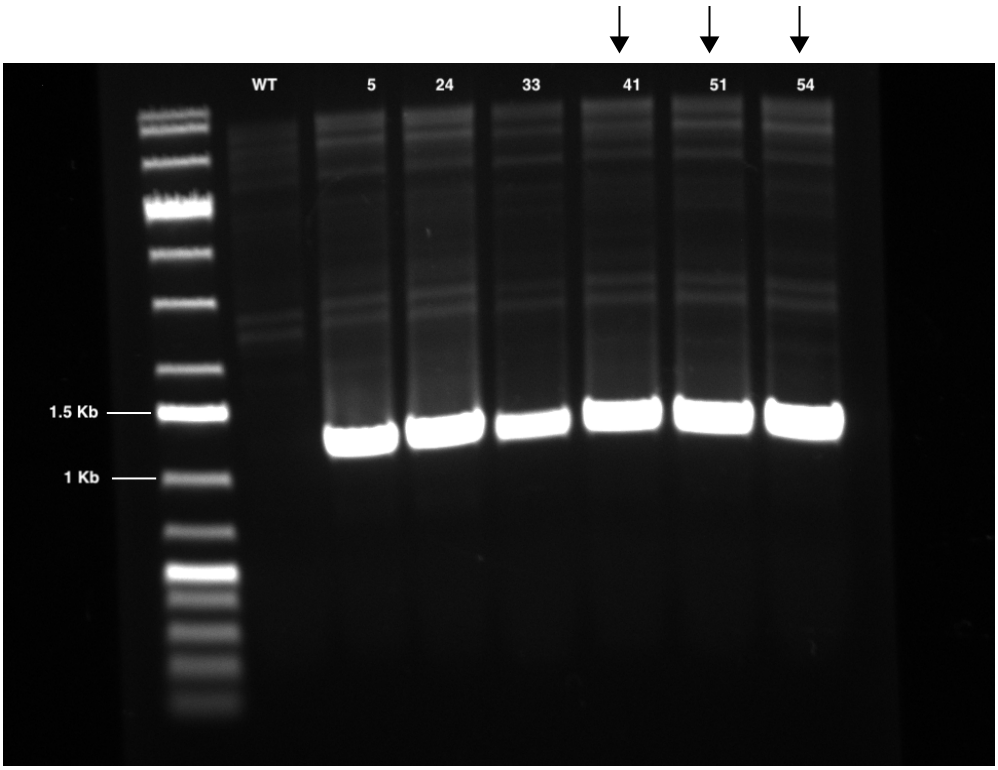

Figure S5: PCR validation of LTR10.KDM6A CRISPR_KO clones. PCR primers used for validation (KDM6A_up_external, KDM6A_down_external) flank LTR10.KDM6A element. Expected KO amplicon size is ~1,245 bp. Expected WT amplicon size (unobserved) is 8,854 bp. Arrows indicate clones that were used for DEseq2 analysis against wildtype HCT116 cells, based on PCA results.

Figure S6A

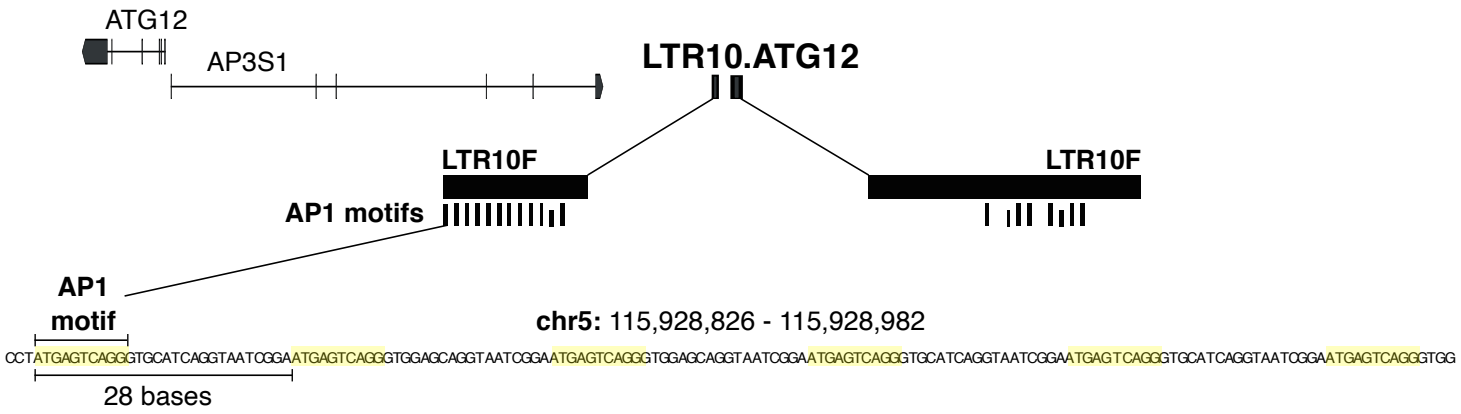

Figure S6A: Genome browser screenshot of LTR10.ATG12 showing a close-up of the 28 bp tandem repeat region that includes the AP1 motif.

Figure S6B

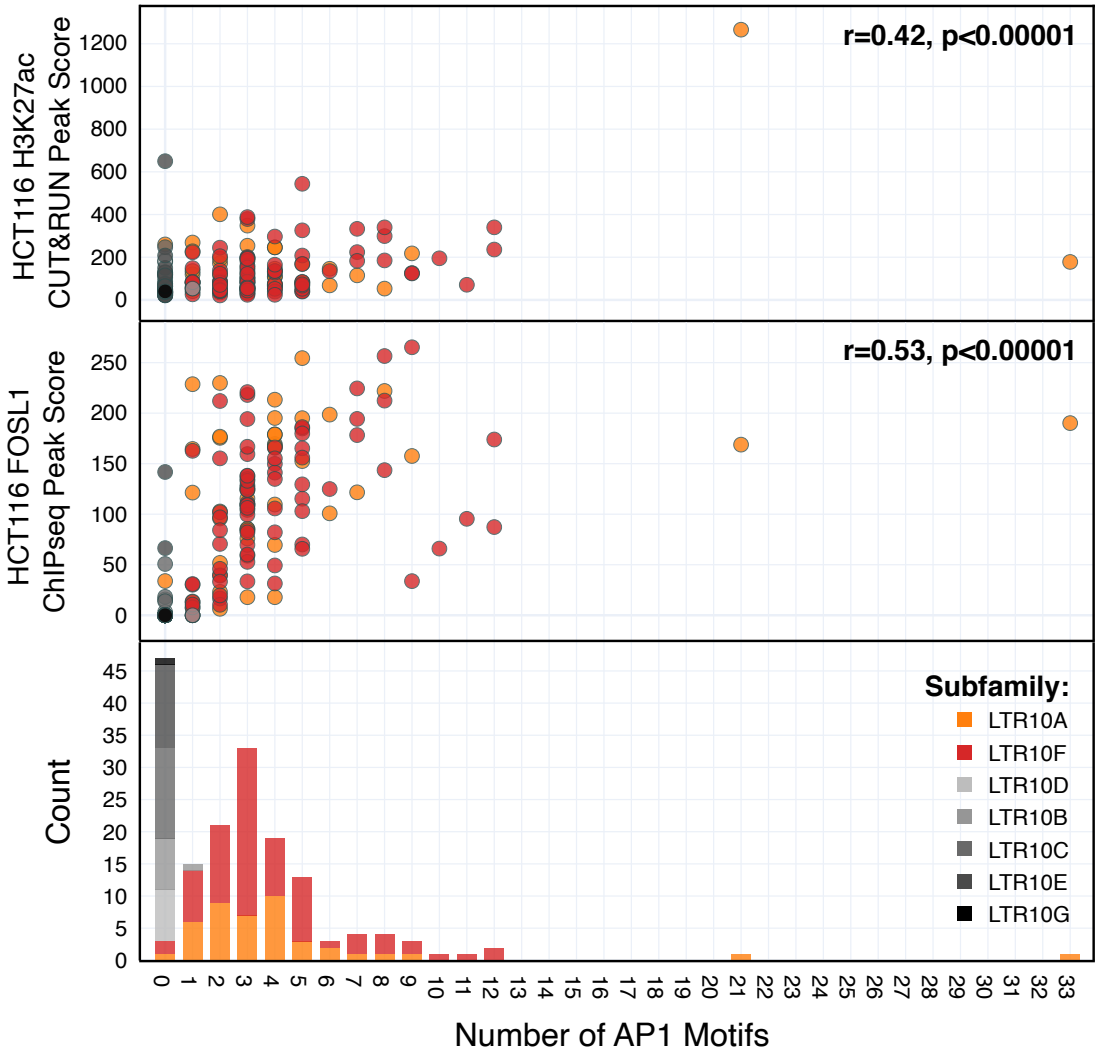

Figure S6B: Scatterplots showing correlations between the number of AP1 motifs with H3K27ac signal (in-house CUT&RUN) and FOSL1 signal (public ChIP-seq, GSE32465). Only LTR10 elements marked by H3K27ac are shown. Enrichment scores were determined using GIGGLE (Methods).

Figure S6C

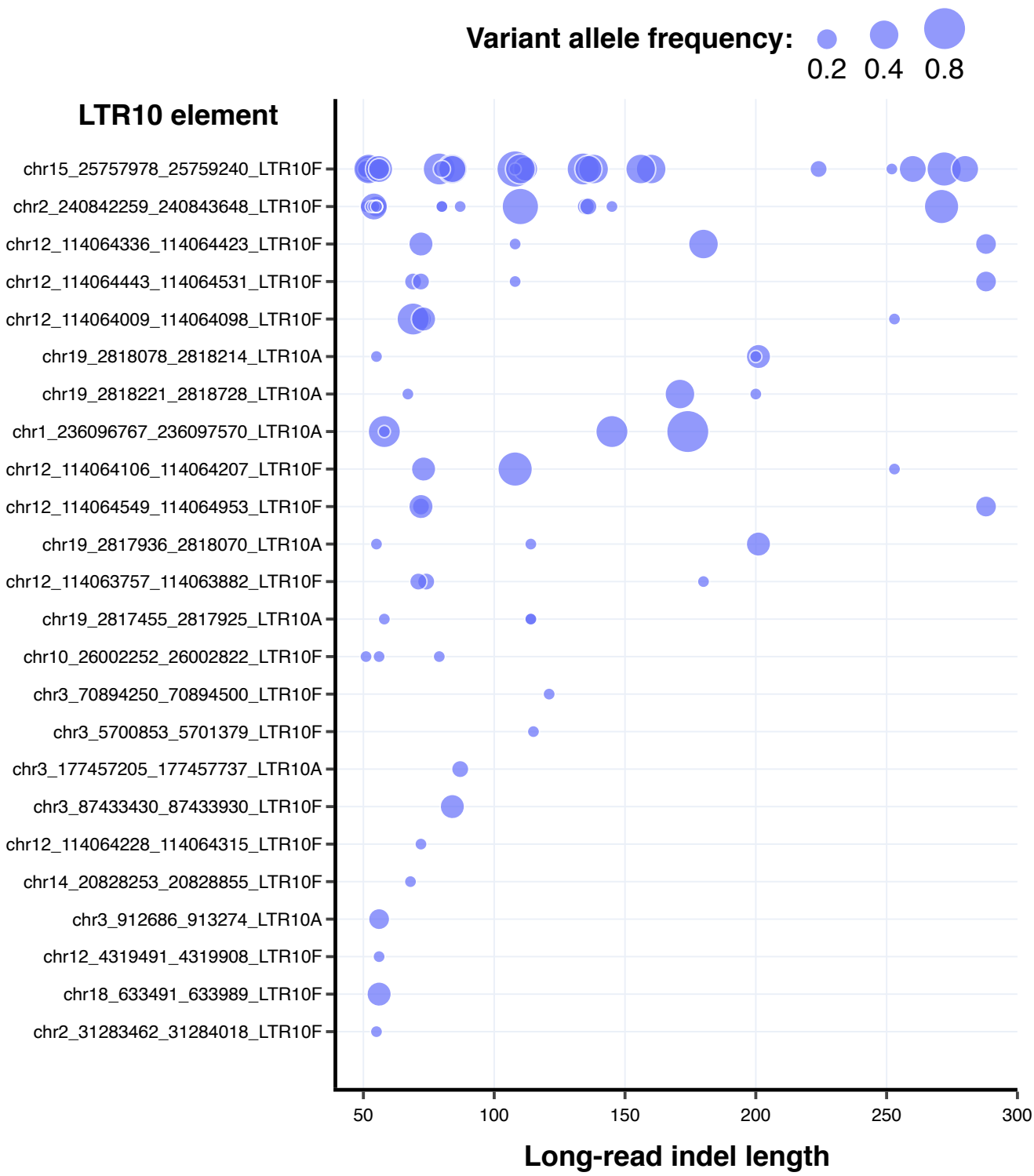

Figure S6C: Scatterplot of long-read indels between 50-300 bp in length overlapping LTR10A or LTR10F elements. Each dot represents a distinct indel, plotted by its length and variant allele frequency (bubble size). Indels were derived from long-read structural variant calls generated from 15 individuals (Audano et al. 2019).

Figure S6D

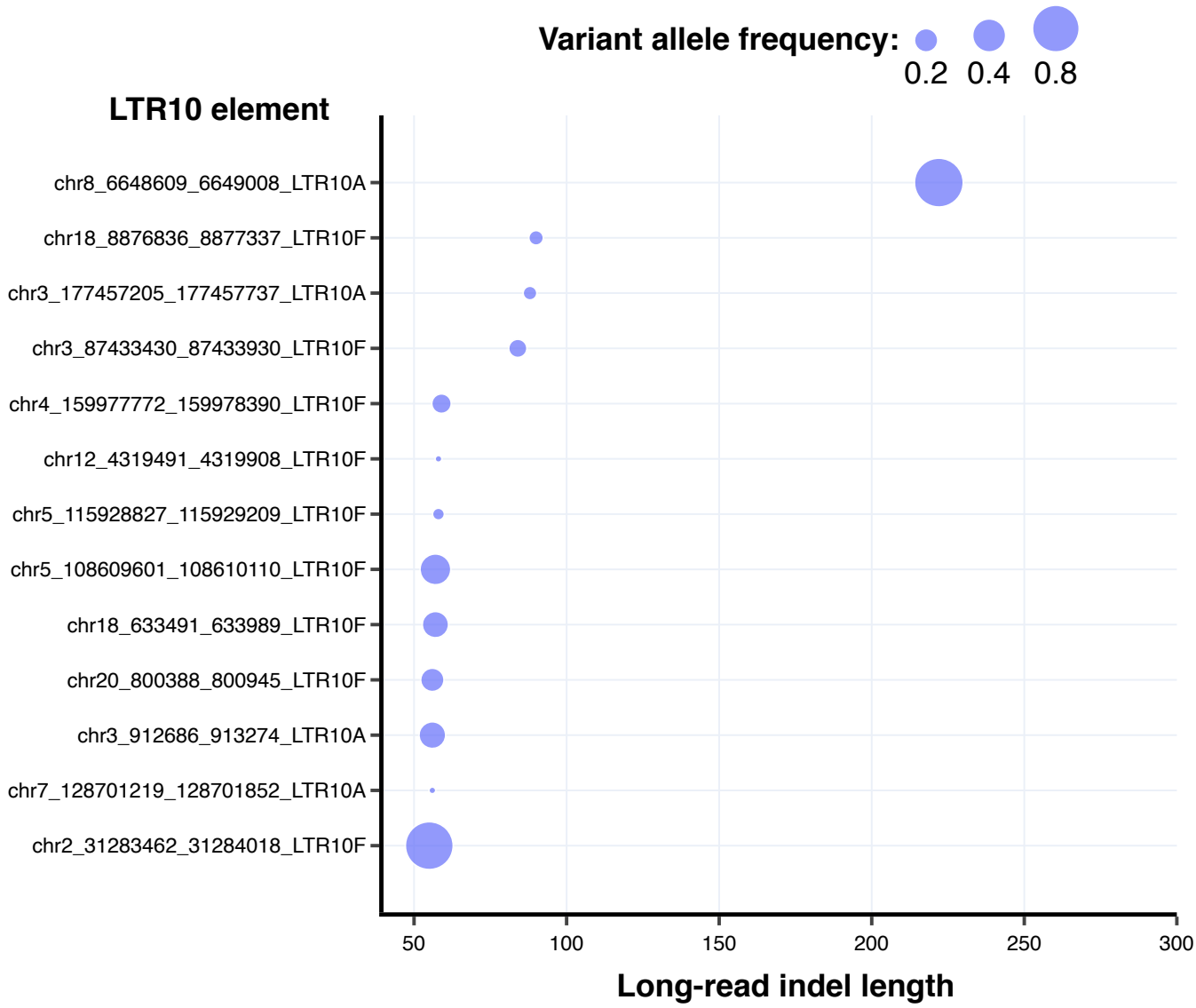

Figure S6D: Scatterplot of long-read indels between 50-300 bp in length overlapping LTR10A or LTR10F elements. Each dot represents a distinct indel, plotted by its length and variant allele frequency (bubble size). Indels were derived from long-read structural variant calls generated from 25 individuals (Quan et al. 2021).

Figure S6E

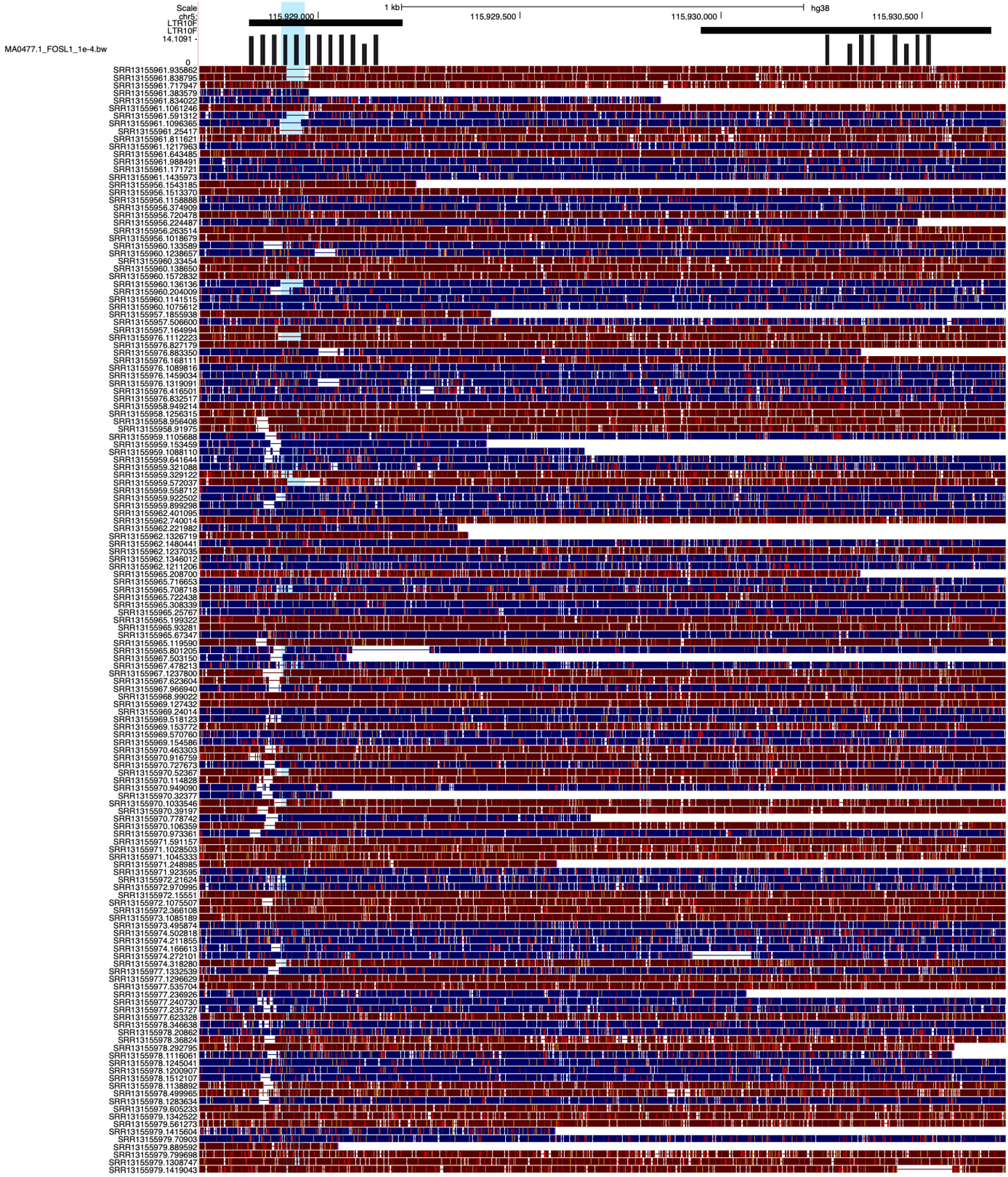

Figure S6E: Genome browser screenshot of a deletion within the LTR10.ATG12 element, showing reads from a long-read dataset of 25 Asian individuals (Quan et al. 2021). The blue highlight shows the location of the reported deletion.

Figure S6F

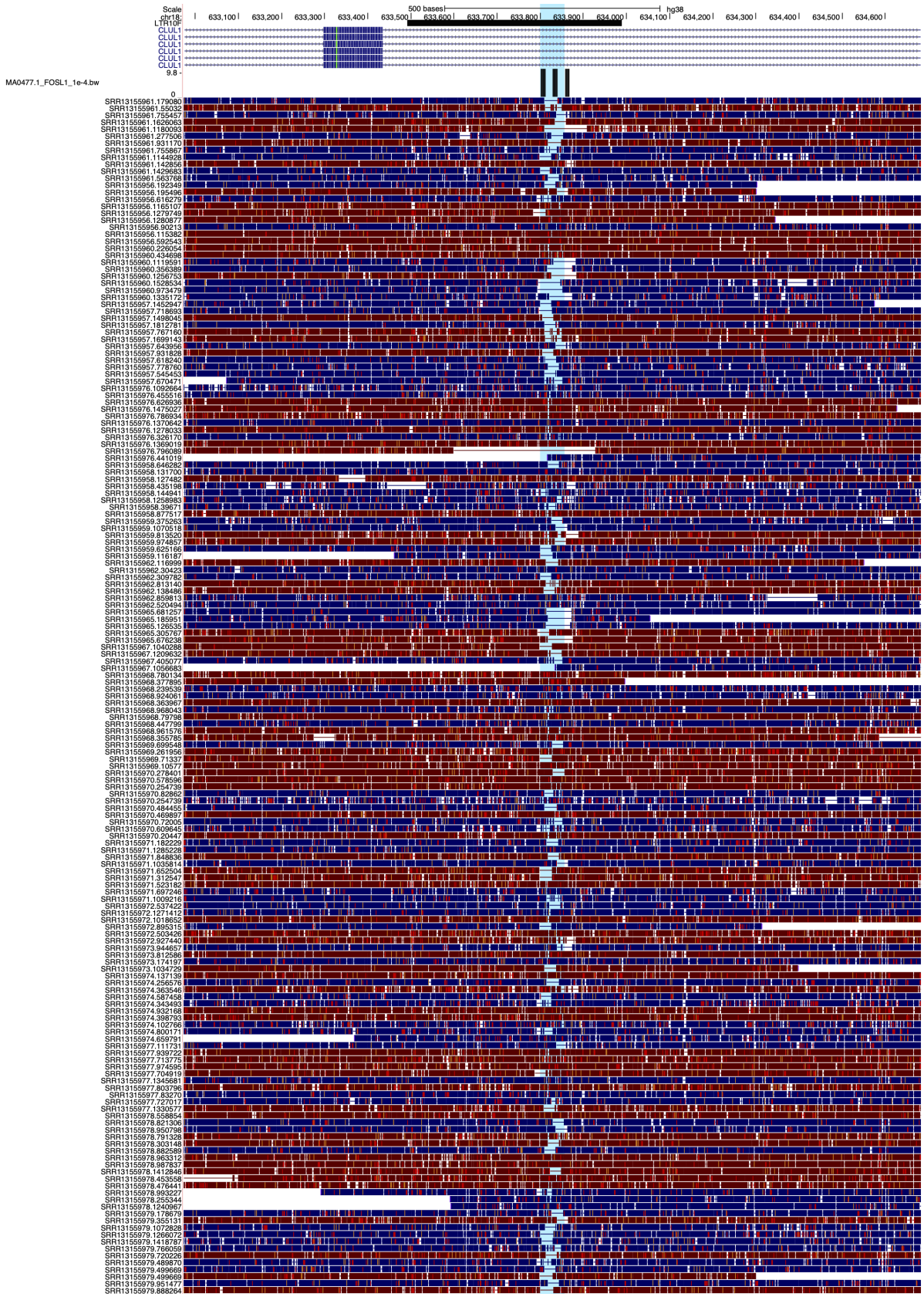

Figure S6F: Genome browser screenshot of a deletion within an LTR10F element within the *CLUL1* gene, showing reads from a long-read dataset of 25 Asian individuals (Quan et al. 2021). The blue highlight shows the location of the reported deletion.

Figure S6G

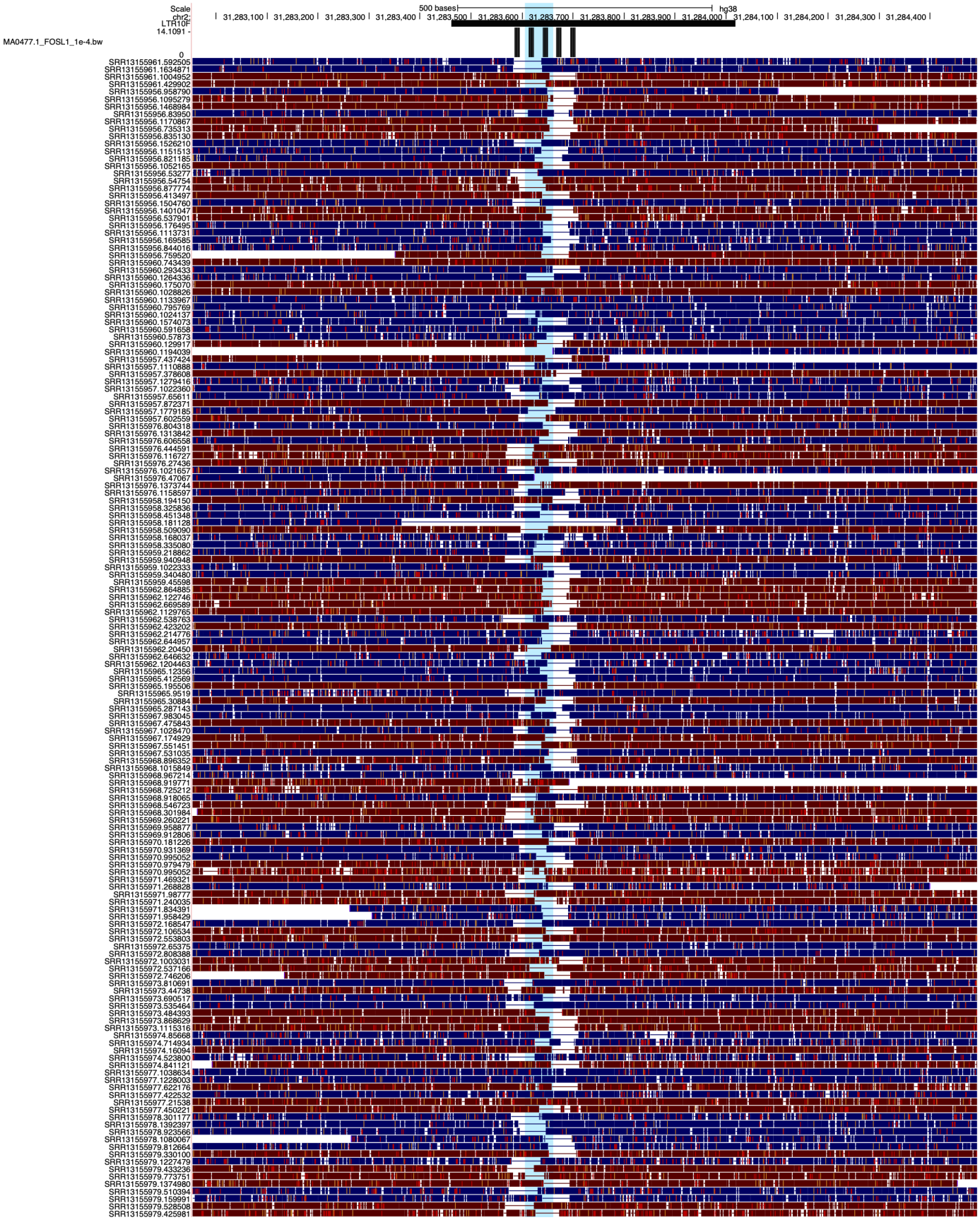

Figure S6G: Genome browser screenshot of a chr2 deletion within an LTR10F element, showing reads from a long-read dataset of 25 Asian individuals (Quan et al. 2021). The blue highlight shows the location of the reported deletion.

Figure S6H

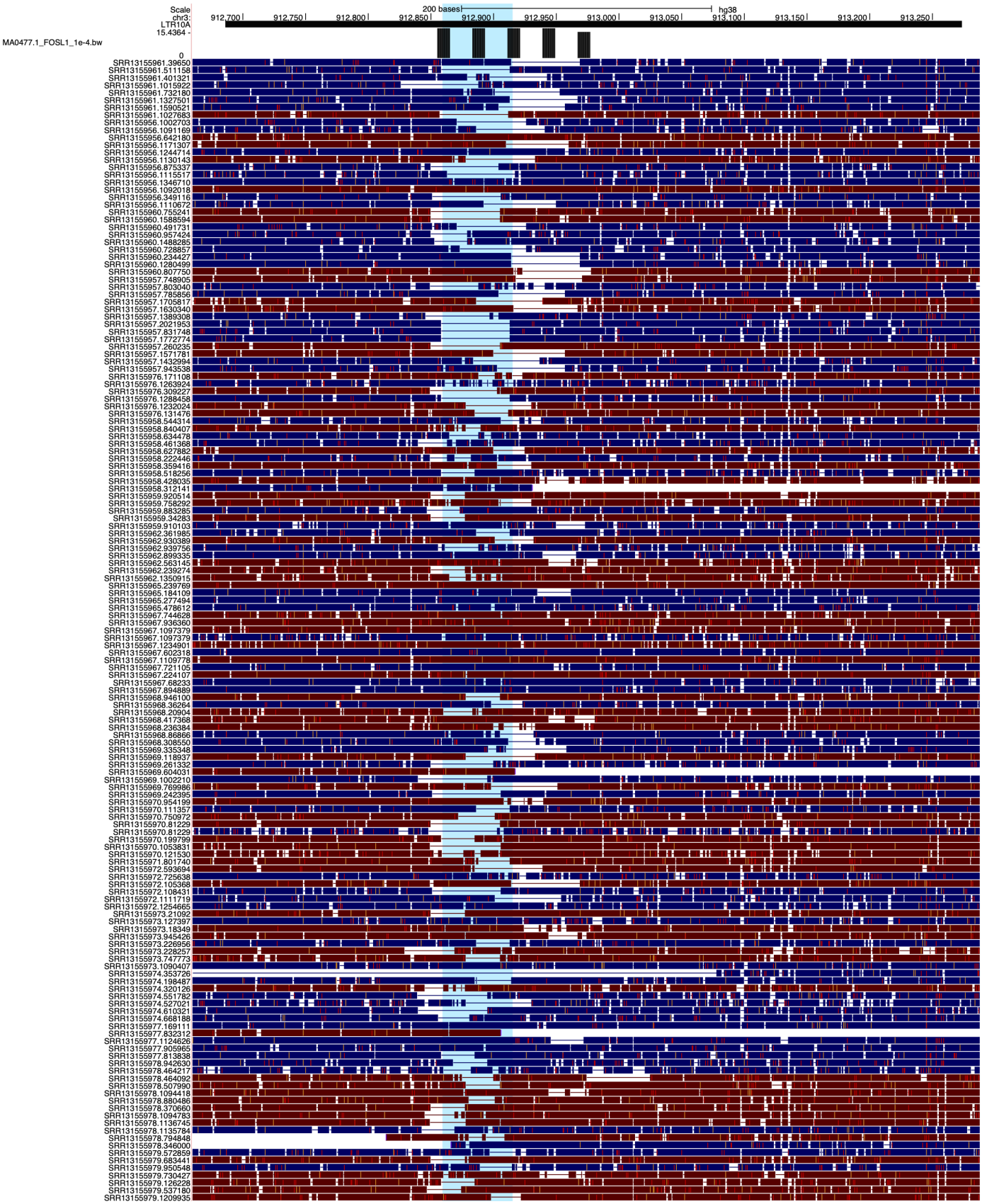

Figure S6H: Genome browser screenshot of a chr3 deletion within an LTR10A element, showing reads from a long-read dataset of 25 Asian individuals (Quan et al. 2021). The blue highlight shows the location of the reported deletion.

Figure S6I

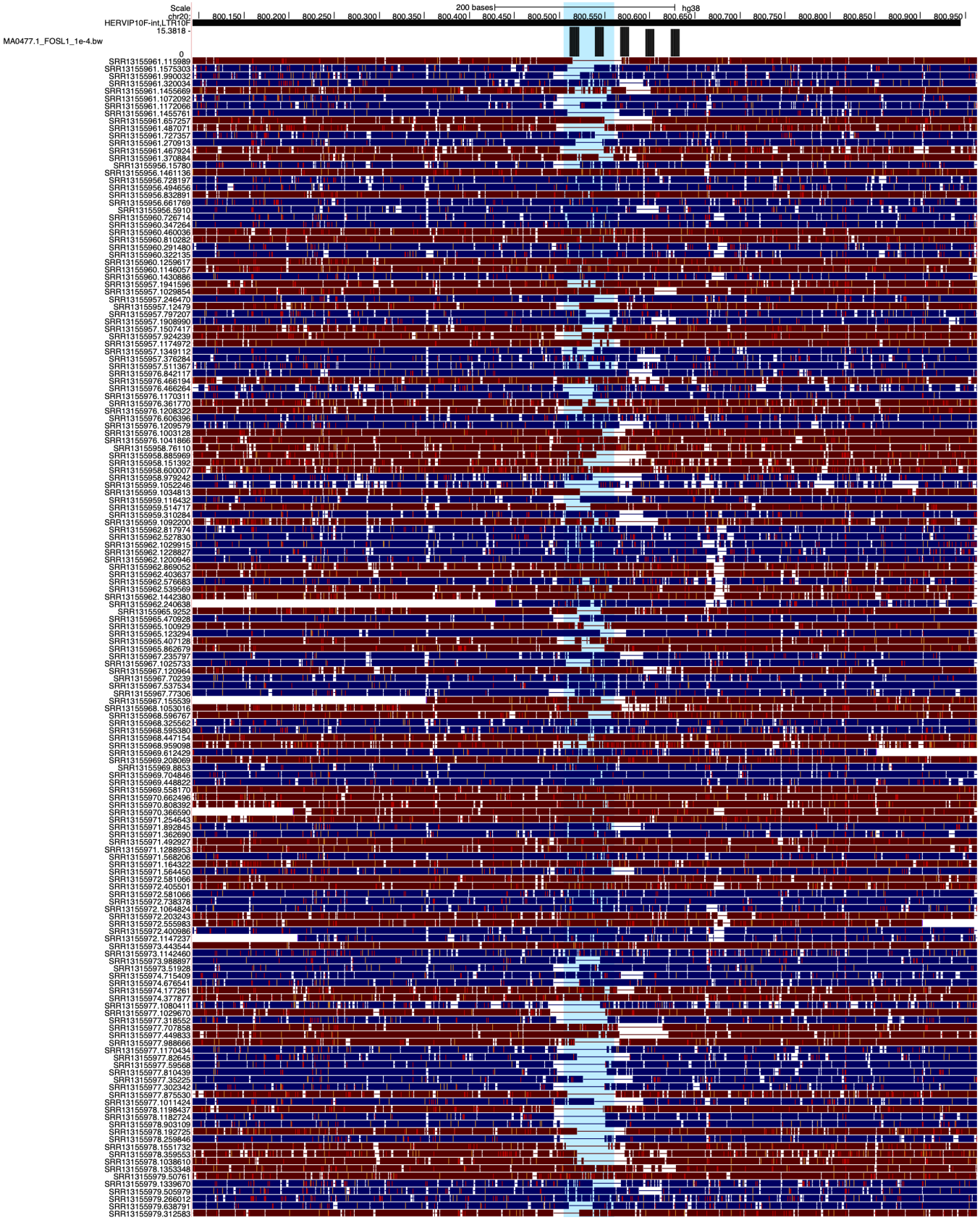

Figure S6I: Genome browser screenshot of a chr20 deletion within an LTR10F element, showing reads from a long-read dataset of 25 Asian individuals (Quan et al. 2021). The blue highlight shows the location of the reported deletion.

Figure S6J

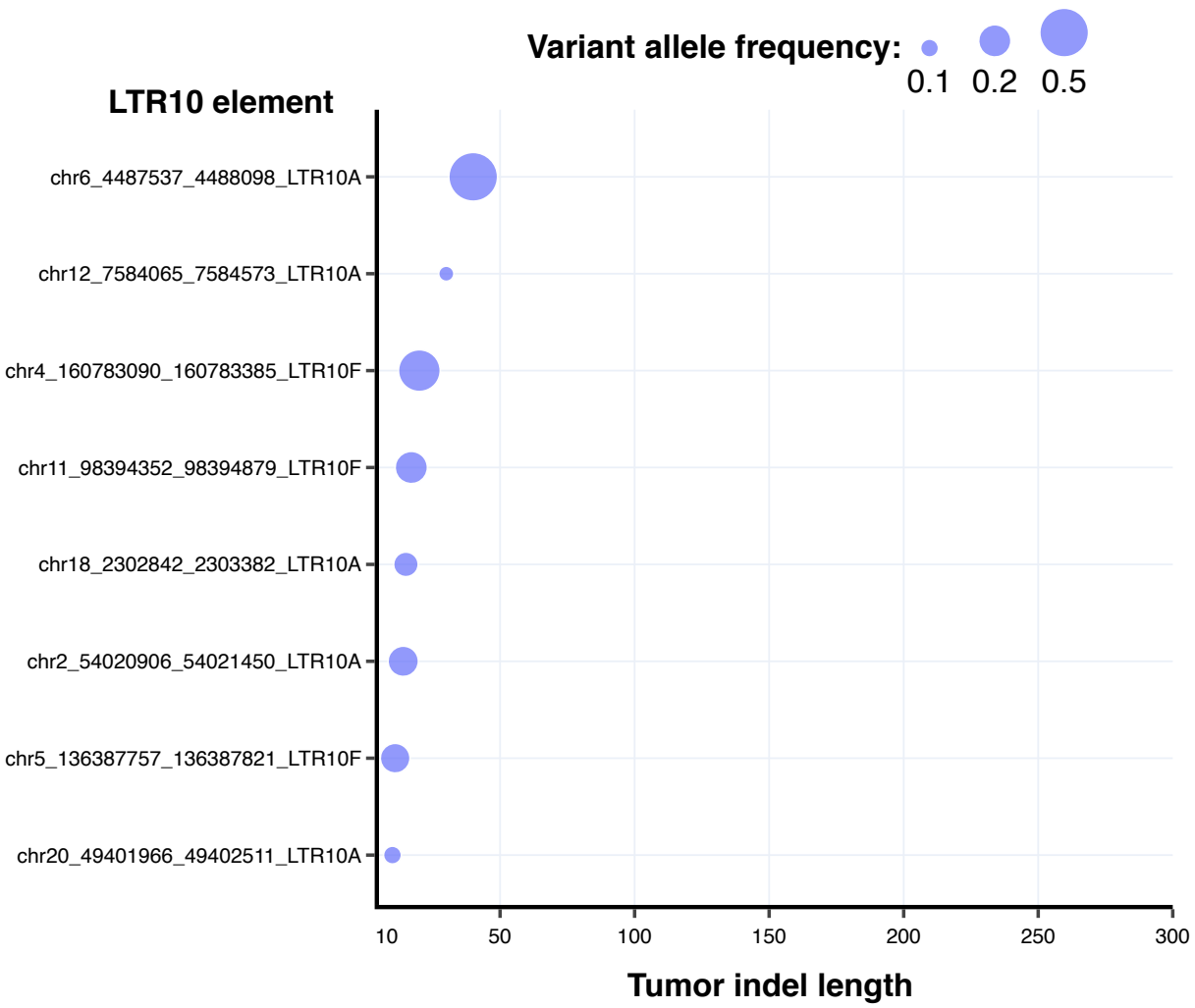

Figure S6J: Scatterplot of somatic indels in patient tumors between 10-300 bp in length overlapping LTR10A or LTR10F elements. Each dot represents a distinct indel, plotted by its length and variant allele frequency (bubble size). Indels were derived from tumor-specific structural variant calls (ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium 2020).