Transcription elongation defects link oncogenic splicing factor mutations to targetable alterations in chromatin landscape

Prajwal C. Boddu[1], Abhishek K. Gupta[1], Rahul Roy[1], Barbara De La Pena Avalos[2 3 4], Ane Olazabal Herrero[1 5], Nils Neuenkirchen[6], Josh Zimmer[7], Namrata Chandhok[8], Darren King[9], Yashuhito Nannya[10], Seishi Ogawa[10], Haifan Lin[6 11], Matthew Simon[7 12], Eloise Dray[2 3 4], Gary Kupfer[5], Amit K. Verma[13], Karla M. Neugebauer[14 15], Manoj M. Pillai[1, 6, 15, 16*]

1. Section of Hematology, Yale Cancer Center and Department of Internal Medicine, Yale University School of Medicine, New Haven, CT, USA
2. Department of Biochemistry and Structural Biology, University of Texas Health Science Center at San Antonio, San Antonio, TX, USA
3. Greehey Children's Cancer Research Institute, University of Texas Health Science Center at San Antonio, San Antonio, TX, USA
4. Mays Cancer Center at UT Health San Antonio MD Anderson, San Antonio, TX, USA
5. Department of Oncology and Pediatrics, Lombardi Comprehensive Cancer Center, GeorgeTown University, Washington DC, USA
6. Yale Stem Cell Center, Yale University School of Medicine, New Haven, CT, USA
7. Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, USA
8. Division of Hematology, Department of Medicine, Sylvester Comprehensive Cancer Center, University of Miami, Florida, USA
9. Section of Hematology and Medical Oncology, Department of Internal Medicine and Rogel Cancer Center, University of Michigan Health, Ann Arbor, Michigan, USA
10. Department of Pathology and Tumor Biology, Kyoto University, Kyoto, Japan
11. Department of Genetics, Yale University School of Medicine, New Haven, CT, USA
12. Institute for Biomolecular Design and Discovery, Yale University, West Haven, CT, USA
13. Division of Hemato-Oncology, Dept. of Medicine and Department of Developmental & Molecular Biology, Albert Einstein-Montefiore Cancer Center, New York, USA
14. Department of Molecular Biophysics and Biochemistry, Yale University School of Medicine, New Haven, CT, USA
15. Yale Center for RNA Science and Medicine, Yale University, New Haven, CT, USA
16. Department of Pathology, Yale University School of Medicine, New Haven, CT, USA

*To whom correspondence should be addressed

Manoj M. Pillai

300 George Street, Suite 786

New Haven CT 06524

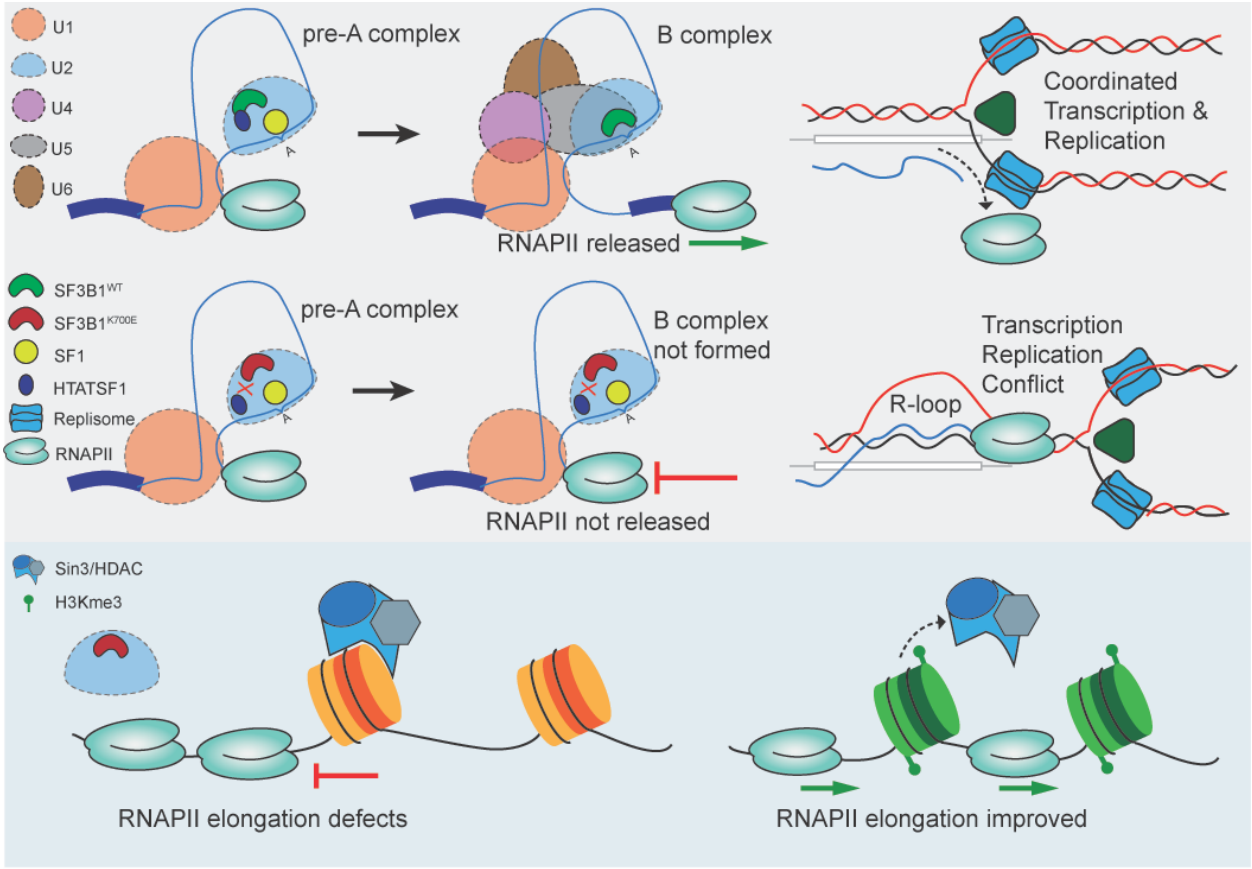
Tel: 203-361-4896

Email: manoj.pillai@yale.edu

HIGHLIGHTS:

- Oncogenic mutations of SF3B1 and U2AF1 cause a gene-body RNAPII transcription elongation defect
- The elongation defect is linked to impaired assembly of early spliceosome complexes and leads to replication stress and changes to chromatin landscape
- RNAPII elongation defects are normalized by modulating epigenetic factors, but alternative spicing events are not reversed
- Targeting the Sin3/HDAC pathway to normalize RNAPII elongation defect is a potential therapeutic approach in these disorders

GRAPHICAL ABSTRACT



ABSTRACT

Transcription and splicing of pre-messenger RNA are closely coordinated, but how this functional coupling is disrupted in human disease remains unexplored. Here, we investigated the impact of non-synonymous mutations in SF3B1 and U2AF1, two splicing factors commonly mutated in cancer, on transcription. We find that the mutations impair RNA Polymerase II (RNAPII) transcription elongation along gene bodies leading to transcription-replication conflicts, replication stress and altered chromatin organization. This elongation defect is linked to disrupted pre-spliceosome assembly due to impaired protein-protein interactions of mutant SF3B1. Through an unbiased screen, we identified epigenetic factors in the Sin3/HDAC complex, which, when modulated, normalize transcription defects and their downstream effects. Our findings shed light on the mechanisms by which oncogenic mutant spliceosomes impact chromatin organization through their effects on RNAPII transcription elongation and present a rationale for targeting the Sin3/HDAC complex as a potential therapeutic strategy.

INTRODUCTION

Although traditionally seen as separate processes driven by distinct multi-protein complexes, we now recognize that RNA transcription and splicing are closely coordinated[1 2]. Accordingly, the spliceosome assembles and removes introns while the nascent transcript is still growing longer, although not strictly in temporal order[3 4]. Perturbing splicing (through inhibitors such as Pladienolide B which inhibits assembly of the U2 snRNP) can affect many aspects of transcription and RNA processing[5 6]. Similarly, slowing the transcriptional machinery changes the selection of alternative splice sites[7]. Despite such close functional coordination, disruption of co-transcriptional splicing as a basis of human disease has not been explored.

Recurrent mutations in select RNA splicing factor (SF) genes, such as *SF3B1*, *U2AF1*, *SRSF2* and *ZRSR2*, are drivers of oncogenesis in many cancers, but are most prevalent in clonal myeloid disorders such as myelodysplastic syndromes (MDS) and acute myeloid leukemia (AML)[8 9 10]. In the decade since their initial description, many insights have been gained into the nature of these mutations and how RNA splicing is altered. SF mutations are usually heterozygous and non-synonymous (except ZRSR2, which are single allele loss-of-function mutations)[11]. They are also largely mutually exclusive, suggesting that shared downstream mechanisms render multiple mutant cells unviable.

SF3B1, the most mutated splicing factor, is a component of U2 snRNP that stabilizes the interaction of U2 snRNA with branch point (BP)[12]. Although highly prevalent in clonal myeloid disorders, SF3B1 mutations are also prevalent in diverse tumor types such as chronic lymphocytic leukemia, uveal melanoma and pancreatic adenocarcinoma[13]. These mutations were largely restricted to the C-terminal HEAT domains (especially HEAT domains 5-9 from a total of 20). The mutations change the SF3B1 confirmation which is believed to reduce BP fidelity, resulting in the selection of cryptic 3'splice sites (SSs)[14]. Altered interactions of mutant SF3B1 with accessory splicing proteins such as SUGP1[15], DDX42, and DDX46[16] are thought to underlie these splicing changes. Such alternative splicing (AS) events, demonstrable in genes such as ABCB7[17], MAP3K7[18], and BRD9[19], are speculated to contribute to different aspects of SF3B1-mutant disease, such as clonal evolution and dyserythropoiesis in MDS. Although AS of cancer-associated genes has been extensively pursued as the primary mechanism for oncogenesis, some caveats remain. Changes to AS, while often demonstrable, are generally modest in degree (low change in isoform ratios or low Percent Spliced-In values) and are not always reproducible across independent models or patient sample datasets[8 19 20]. Importantly, patterns of AS changes common to different SF mutations have not yet emerged. These, and the observation that SF mutations are largely mutually exclusive, has prompted investigation into alternate mechanisms that do not rely on mutant-specific changes in isoform abundance.

In this study, we explored the mechanisms by which mutant spliceosomes dysregulate transcription kinetics. We show that mutations in SF3B1 and U2AF1 primarily dysregulate RNAPII gene-body elongation due to altered assembly of early spliceosomes, leading to replication stress and chromatin re-organization. Altered RNAPII elongation was corrected by modulation of specific epigenetic pathways, underscoring the interplay between dysregulated transcription and chromatin organization, and identifies targets of therapeutic relevance.

RESULTS

SF3B1^{K700E} redistributes RNAPII from promoters to gene bodies

To explore the functional and pathologic relevance of disruption of co-transcriptional splicing in SF mutant disease context, we developed and optimized a novel inducible isogenic cell line system combining AAV-intron trap and CRISPR/Cas9 genome editing that allows for expansion of cells upfront and expression of mutant allele upon inducible expression of Cre-recombinanse[21] (Fig1A). We generated inducible isogenic cell lines for the K700E and S34F mutations in *SF3B1* and *U2AF1*, respectively- the most common cancer associated mutations in each of those genes. Control cells were edited with AAV-intron traps with synonymous mutations. To determine changes to transcription by SF3B1^{K700E}, we started with ChIP-seq for RNAPII utilizing a pan-RNAPII antibody (NTD) that recognizes RNAPII irrespective of its activation status (FigS1A). Metaplot of average RNAPII occupancy across 6694 highly expressed genes in K562 cells (determined from RNA-seq per Ciazzi et al.[5] and validated using in our RNA-seq datasets) revealed a redistribution of RNAPII in SF3B1^{K700E} into gene bodies (Fig1B). RNAPII density over the promoter/transcription start site (TSS) was reduced while density across the gene body progressively increased culminating at the TTS. Representative density plots of such redistribution are shown in Fig1C. The change was also captured by the traveling ratio (TR)[22 23], which compares read density in the promoter region to that in gene body (Fig1D). ChIP-seq with an elongation-specific RNAPII antibody (RNAPII Ser2P)(FigS1B) also showed a marked redistribution of RNAPII occupancy, suggesting a primary failure of RNAPII elongation in SF3B1^{K700E} cells (Fig1E, FigS1C). To determine how gene-specific features influence RNAPII elongation defects in SF3B1^{K700E} cells, we performed unsupervised clustering to divide the genes into three clusters, based on the changes in TR (delta TR) (Fig1F, FigS1D). We then determined multiple attributes of the genes within these individual clusters. Notably, the clusters differed significantly with respect to introns per gene and total gene length in a consistent trend (genes that were longer or contained more introns had more RNAPII redistribution) (Fig1G). Other parameter such as gene expression levels and distance to neighboring genes did not follow a consistent trend across the clusters (FigS1E). Notably, the cluster with least change in TR (Cluster 3) was significantly enriched for intronless genes, containing over half of the intronless genes (102/200) in the dataset (Fig1H). In an alternative analysis, we directly binned genes based on their length or intron number (FigS1F). Delta TR independently correlated with gene length and intron number (FigS1G). We also analyzed the RNAPII NTD ChIP data to determine how the transcription elongation behavior changes across the gene body with static binning (instead of dynamic binning used in gene-body length normalization in metagene plots). As shown in Fig1I, SF3B1^{K700E} cells showed a relative decrease in RNAPII density downstream of TSS followed by a sharp increase in the immediate gene bodies (note the red-to-blue region after the TSS in the heatmap; Fig1I). This transition zone occurred before the (TSS+0.5kb) region, coinciding with the first exon-intron junction boundary (Fig1J). To investigate the possibility of whether RNAPII enrichment was selective

to the location of the intron in the gene body, RNAPII ChIP-seq densities for the first, second, and terminal introns were independently analyzed and noted to show a similar profile (FigS1H). Together, we conclude that RNAPII in SF3B1^{K700E} cells redistribute from promoters to gene bodies, dependent on the cumulative distance RNAPII has to traverse (gene length) and the number of spliceosomes assembled in this distance (total introns).

RNAPII redistribution in SF3B1^{K700E} results from changes in gene body elongation rate

Redistribution of RNAPII in SF3B1^{K700E} can arise from two possible scenarios: (a) increased productive elongation resulting from increased transcription initiation and/or premature pause release that reduces pileup of RNAPII at promoters, or (b) defective elongation of RNAPII along the gene body. Since RNAPII ChIP-seq cannot distinguish between the two scenarios, we implemented two complementary transcriptomic strategies to determine how the redistributed RNAPII transcribes nascent RNA (fastGRO-seq and TT-TL-seq). FastGRO-seq (referred to as GRO-seq now on)[24] is a nuclear run-on assay based on the classical GRO-seq protocol that uses 4-thio ribonucleotides (4sUTPs) to label nascent RNA arising from transcriptionally engaged RNAPII. High correlation between replicates indicates the reproducibility of GRO-seq (FigS4A). Aligned reads were split to sense (+) or antisense (-) and metagene plots were generated. As previously reported, sense read profiles had with a sharp peak at TSS, then tapering through gene body with a second peak at TTS; Antisense reads were largely 5' of the TSS (Fig2A). Metaplots revealed similar promoter peak densities in SF3B1^{K700E} and SF3B1^{WT}, but a higher gene-body signal in SF3B1^{K700E} peaking at the TTS (Fig2A-B). This was also reflected in the TR calculated from GRO-seq signals (Fig2C). The GRO-seq thus confirms our ChIP-seq results that RNAPII in SF3B1^{K700E} redistributes into the gene body, which when released during run-on transcription results in higher gene body density compared to SF3B1^{WT}. TR calculation analyses of GRO-seq signal were further broken down by cluster I, II, and III genes (as in Fig1F), with groups I and II exhibiting a more significant increase of nascent RNA within gene bodies (FigS2B). Consistent with the preferential accumulation of RNAPII over introns in RNAPII ChIP seq, elevated GRO-seq signal at introns was observed (Fig2D).

Enhanced pause release of RNAPII and reduced RNAPII elongation rate may be distinguished by comparing the relationship between steady state transcript and nascent transcription levels[25 26]. Upregulated nascent RNA synthesis will be reflected in the steady state transcript levels in the former. Conversely, there will be little to no correlation between nascent RNA and steady RNA profiles in the latter scenario. Read densities across gene bodies from GRO-seq and corresponding total RNA-seq were quantified and compared using the DESeq2 package. As shown in Fig2E, those genes with differential steady state levels (total RNA-seq in SF3B1^{K700E} vs SF3B1^{WT}) did not correlate with their corresponding GRO-seq values. As complementary approaches to GRO-seq, we additionally performed TT-TL-seq[27] and nascent short-read (nSRS) RNA-seq[28] in WT and SF3B1^{K700E} (Fig2F, S2C-D). TT-TL-seq combines transient transcriptome sequencing (TT-seq, which labels and captures

newly synthesized transient RNA) with "time lapse" chemistry (converts 4sU to cytidine analogs, yielding apparent U-to-C mutations that can identify new transcripts in the next generation sequencing datasets (FigS2E). The time lapse component allows precise identification of nascent RNA reads based on the U-to-C conversion (Fig2F). Metagene analysis showed that the Drosophila-spike-in normalized[29] TT-seq coverage across the gene body was reduced in SF3B1^{K700E} (Fig2G), consistent with reduced RNAPII elongation rate as opposed to enhanced/premature pause release. Differential gene expression analysis of nascent transcription profiles based on a gene body window 300 bp downstream of TSS showed 48.7% and 1.3% of genes with significant decreased and increased nascent RNA transcription levels, respectively (FigS2F). Notably, the degree of reduction of the TT gene body signal did not correlate with baseline gene expression, a pattern consistent with the lack of a direct physical effect of SF3B1^{K700E} on the transcribing RNAPII[5] (FigS2G). Finally, hierarchical clustering of gene expression changes on the four transcriptomic datasets (total RNA-seq, nSRS-seq, GRO-seq and TT-TL-seq) showed total RNA-seq profile to cluster differently from the 3 nascent RNA-seq measures (Fig2H). Collectively, our data shows that RNAPII redistribution results from impaired elongation and not premature pause release.

Given the mutual exclusivity of SF mutations, we speculated if a similar RNAPII transcriptional response is demonstrable in U2AF mutant cells. Using an inducible *U2AF1* ^{S34F} isogenic cell line system[21], we similarly observed redistribution of RNAPII into gene bodies by Ser2P RNAPII ChIP (FigS2H), and a globally reduced gene-body signal by TT-TL seq (FigS2I). These data suggest a convergent transcriptional response to distinct SF mutations, which may underlie their mutual exclusivity.

Long-read sequencing (LRS) reveals reduced co-transcriptional splicing in SF3B1^{K700E}

Given our findings that RNAPII elongation rate is reduced by mutant SFs, we wondered how co-transcriptional splicing is affected by the mutations. The effect of SF mutations on co-transcriptional splicing efficiency (CoSE) itself has not been explored. Importantly, a common pattern of alternative splicing noted across mutant SF patient samples is a decrease in retained introns (RI)[30]. It is currently unclear if this reduced RI results from improved efficiency of co-transcriptional splicing. An alternative possibility is that such introns are removed from transcripts post-transcriptionally. We have previously reported on the use of LRS on the PacBio platform to determine kinetics of co-transcriptional splicing[1]. LRS can give single molecule-level metrics which Illumina short read sequencing cannot provide. PolyA-depleted chromatin-associated RNA (nascent RNA) was isolated as previously described[28] for PacBio library preparation. The libraries had mean read lengths between 728 and 901 nucleotides, and 85-90 average reads per gene (FigS2J-K). A representative example of aligned LRS reads (ordered by 3' end position and colored based on splicing status) is shown, in FigS2L. The 3' end of each LRS read signals the position of RNAPII and could be used to determine distribution of RNAPII across the gene body. Plotting the median RNAPII position across normalized gene length as a cumulative distribution function showed a lag (shift towards 5' position) in median RNAPII

positions in SF3B1^{K700E} (Fig2I). We then divided the reads to the 3 gene clusters (as determined by RNAPII ChIP-seq data, Fig1F) and determined the change in 3' position in each of the clusters. As seen in GRO-seq and ChIP-seq datasets, relative lag of median RNAPII position in SF3B1 ^{K700E} was most pronounced in clusters 1 and 2 (FigS2M). LRS thus provides an additional line of evidence for slower RNAPII gene-body elongation in SF3B1^{K700E}.

We then used LRS to determine CoSE for individual introns at a single molecule resolution. CoSE for each splice junction was calculated as a ratio of spliced transcripts to total transcripts (spliced and unspliced) that traverse that junction. A lower CoSE would indicate a lower rate of co-transcriptional splicing at that junction and vice-versa. Genome-wide comparison of splice junctions showed a significantly reduced CoSE in SF3B1^{K700E}, compared to SF3B1^{WT}(Fig2J). We then determined if the position of introns with relation to transcription start and end sites influenced changes to CoSE. To this end, the CoSEs for the first, middle, and final intron subgroups were separately analyzed and showed more pronounced reduction in CoSEs in middle and final intron subgroups (FigS2N). We also leveraged our GRO-seq datasets for the occurrence of reads, separately for the 5' splice sites and 3' splice sites and observed a statistically significant reduction in splicing ratios in SF3B1 ^{K700E}, both, when all introns were considered, as well as when first, middle, and last introns were analyzed (Fig2K, FigS2O). Finally, we specifically compared the CoSE of those introns, with reduced IR (identified from RNA-seq using IRFinder[31]), in isogenic SF3B1^{K700E} cells. Interestingly, these introns also had a reduced CoSE in SF3B1^{K700E} compared to SF3B1^{WT} (Fig2L, FigS2P). Together, these results suggest co-transcriptional splicing is not increased in SF3B1^{K700E} and that the reduced IR observed in a subset of genes across MDS patient datasets is unlikely to be a result of more efficient co-transcriptional splicing.

SF3B1^{K700E} disrupts E-to-A spliceosome complex transition

Transcription is regulated by a complex array of initiation/elongation complexes. A change in expression levels of such factors (by AS or other indirect mechanisms) may be invoked to explain RNAPII elongation defects[32]. Analysis of RNA-seq datasets of our isogenic cells, or public databases[33 34 35] did not show reproducible changes to AS or gene expression levels of transcriptional regulators. Hence, we explored direct interactions between spliceosome and RNAPII to explain the observed defects in transcription elongation. Such direct interaction between RNAPII and the early spliceosome complex (specifically U1 snRNP) was recently proposed to underlie functional coupling of transcription and splicing[2]. Accordingly, RNAPII is released from this coupling only upon formation of the spliceosome B complex. Recent cryoEM studies have also identified the critical initial step in the pre-spliceosome formation (pre-A or E complex) to be the displacement of HTATSF1 from SF3B1 by DDX46[36,37]. This ATP-dependent step facilitates displacement of splicing factor-1 (SF1) from the BP, followed by binding of U2AF2 to SF3B1 (UHM to ULM domain interaction), and stable binding of U2 snRNP complex to the BP. HEAT domain mutations in SF3B1 (or

Hsh155) are known to impair its interaction with DDX46 (or Prp5)[38 16]. We hence speculated if defective early spliceosome assembly is responsible for RNAPII elongation defects in SF3B1^{K700E}. To explore this, we transfected FLAG-SF3B1 (^{K700E} and ^{WT}) into HEK293T cells, and immunoprecipitated it from nuclear lysates after 48 hours. As shown in Fig3A, FLAG-SF3B1^{K700E} from nuclear lysates showed reduced binding with HTATSF1, U2AF1, and U2AF2. Since spliceosomes are assembled on chromatin[39 40], these interactions were also probed on chromatin fractions and the differences were noted to be even more pronounced (Fig3A). Of note, the input fraction of chromatin also revealed lower chromatin abundance of HTATSF1 in FLAG-SF3B1^{K700E} (vs FLAG-SF3B1^{WT}) implying a role for SF3B1 in engaging HTATSF1 onto chromatin (Fig3A). We next sought to determine specific steps in spliceosome assembly that may be perturbed by these defective interactions. Nuclear lysates from isogenic K562 cells (SF3B1^{K700E} and SF3B1^{WT}) were further divided to nucleoplasmic and chromatin fractions and probed for abundance of splicing proteins, including those specific to E (SF3A3, HTATSF1) and B (MFAP1) complex assemblies (Fig3B). As anticipated, chromatin abundance of HTATSF1 was reduced in SF3B1^{K700E} cells. SF3A3 chromatin levels (E-complex component) were unchanged but MFAP1 (B complex component) was reduced in SF3B1^{K700E} cells, suggestive of a defective transition to B complex. Additionally, we observed higher chromatin levels of SF1 and U2AF2 in SF3B1^{K700E} cells, implying reduced B-complex spliceosome intermediates. We then determined how assembly kinetics of SF3B1^{K700E} differed from SF3B1^{WT} by *in vitro* splicing assay. Nuclear extracts from HEK293T cells transfected with SF3B1^{WT} or SF3B1^{K700E} were used for splicing reactions at varying time points and resolved on 1.5% native agarose gels. Transition of E to A complexes in each of the conditions by determining the A/E ratio at each time point (Fig3C). Transition from E- to A- complex intermediates was reduced in SF3B1^{K700E} nuclear extracts compared to SF3B1^{WT}, most apparent at 40 minutes. Given these results, we speculated if disruption of the early pre-spliceosome assembly (through loss of DDX46 or HTATSF1) would result in a transcriptional response similar to that seen with SF3B1^{K700E}. We used stable inducible lentiviral shRNA to knock down DDX46 and HTATSF1 in wild-type K562 cells (FigS3A) and performed Ser2P RNAPII ChIP-seq. As shown in Fig3D, RNAPII was redistributed into the gene bodies upon loss of either DDX46 or HTATSF1, in a pattern similar to that with SF3B1^{K700E}. These results were also confirmed using transient siRNA-mediated knockdown of DDX46 and HTATSF1 in K562 cells (FigS3B-C). Finally, we determined if growth arrest of SF3B1^{K700E} cells could be rescued by overexpression of either DDX46 or HTATSF1 (FigS3D). Using doxycycline inducible constructs (FigS3D), we found that overexpression of HTATSF1, but not DDX46, led to partial rescue of SF3B1^{K700E} cells (Fig3E). Collectively, these findings are consistent with a model of impaired interaction of SF3B1^{K700E} with HTATSF1, which prevents efficient assembly of spliceosome complex B (Fig3F) which leads to defective RNAPII elongation.

Effect of SF3B1^{K700E} on RNAPII kinetics is distinct from SF3B1 inhibition

Disease associated mutations of SF3B1 are thought to be neomorphs and not loss of function variants. This is based on the nature of mutations (non-synonymous and never loss of function), and selection of novel 3'SSs[30]

[14]. We sought to determine how SF3B1^{K700E} compares to SF3B1 chemical inhibition, in its effects on transcription. Previous studies have shown that acute inhibition of SF3B1 with PlaB largely results in RNAPII proximal-promoter (p-p)[5 41] pausing, a pattern distinct from the RNAPII redistribution in response to SF3B1^{K700E}. SF3B1 inhibition is also known to impair p-p recruitment of elongation complex p-TEFb to RNAPII which then facilitates release of RNAPII from a paused state at the promoter[42 43]. We therefore performed ChIP-seq for CDK9 (a component of the p-TEFb complex) in SF3B1^{K700E} and SF3B1^{WT} cells (Fig3G, FigS3E). SF3B1^{K700E} expression resulted in increased CDK9 p-p occupancy, suggesting that p-TEFb p-p recruitment is not impaired. Chemical inhibitors of SF3B1 such as Pladienolide B (PlaB) are acutely toxic to cells, hence it has been suggested that at least some of PlaB-induced transcriptional alterations are indirect due to cellular stress response[44]. Transcriptional changes in response to cellular stresses include read-through transcription downstream of termination sites (also called Downstream of Genes or DoGs)[41 45]. We utilized data from LRS, GRO-seq and TT-TL-seq to determine extent of readthrough transcription in SF3B1^{K700E} cells. We did not find evidence of increased readthrough past the TTS (Fig3H, FigS3F-G) with any of the three approaches, arguing against perturbation of cleavage/ termination in SF3B1^{K700E}. Finally, pathway analysis of 4-day RNA-seq datasets from these cells do not show an upregulation of stress-related pathways in SF3B1^{K700E} (FigS3H). Collectively, these findings indicate that SF3B1^{K700E} -induced transcriptional changes are distinct from stress-related transcriptional disruption.

SF3B1^{K700E} induced transcription dysregulation is associated with R-loops, transcription replication conflicts (TRCs) and S-phase arrest

While these mutations confer growth advantage *in vivo*, they are known to paradoxically slow growth in cell lines[46,47,48,9]. Consistent with this, isogenic SF3B1^{K700E} cells grow slower and eventually fully growth-arrest in about 2 weeks after mutant allele expression (FigS4A). SF-mutant cells have been shown to have increased levels of R-loops and DNA damage response[46 49]. We reasoned that the dysregulated RNAPII transcription may be driving R-loops resulting in DNA damage and cell growth arrest. We checked for increased R-loop levels in SF3B1^{K700E} using S9.6 monoclonal antibody (Fig4A). S9.6 is known to also bind to dsRNA and ribosomal RNA[50], especially in nucleoli, which can confound R-loop staining. We therefore used nucleolin antibody staining to identify areas of nucleolar signal and subtracted those regions from the nuclei to obtain S9.6 signal in surrounding chromatin regions, as previously described[51]. Dot-blot assay using S9.6 on chromatin preps corroborated our microscopy results (FigS4B). Unresolved R-loops pose a major obstacle to replication fork (RF) progression, by colliding with the DNA replisome machinery, leading to TRCs. To determine if transcription and replication machineries are in close proximity, we performed in-situ proximity ligation assay (PLA)[52] using antibodies against RNAPII and PCNA (protein component of replication complex). As shown in Fig4B, SF3B1^{K700E} nuclei were noted to have significantly higher foci of co-localized transcription and replication, characteristic of TRCs. Further characterization revealed SF3B1^{K700E} to induce

DNA damage, as determined by γH2AX Western blotting (FigS4C) and neutral comet single cell gel electrophoresis assays (FigS4D). Increased replication stress in SF3B1^{K700E} was additionally confirmed using more specific markers, including IF staining for (1) increased ssDNA (BrdU labeling followed by anti-BrdU antibody[53] IF) (FigS4E), and (2) phosphorylated RPA2 ser33 (pRPA)[54] (FigS4F). BrdU and pRPA signals, as well as S9.6 staining, were specifically increased in those cells with increased γH2AX staining (FigS4E-G), indicating that R-loop related DDR is associated with increased replication stress. To further define mechanisms by which SF3B1^{K700E} perturbs R-loop homeostasis, we depleted representative factors from disparate pathways regulating R-loop homeostasis with siRNA: UAP56 (prevents nascent RNA back-hybridization[55]), Senataxin (resolves R-loops by promoting transcription termination[56]), and FANCD2 (a DDR factor[57])[58] (FigS4H). As anticipated, depletion of each of these factors increased both TRCs and DNA breaks (FigS4I) in the WT cells. In SF3B1^{K700E} cells, depletion of Senataxin alone (but not UAP56 or FANCD2) was associated with a further increase in both TRCs as well as DNA breaks (Fig4C), compared to control siRNA. Indeed, overexpressing bacterial RNaseH1 in SF3B1^{K700E} (FigS4J), improved cell growth, decreased comet tail moments, γH2AX foci and R-loops confirming R-loops to be a major contributor to observed DDR (Fig4A-B,4D; FigS4B-D). Given the role of Senataxin and RNaseH1 in directly resolving R-loops, these results further confirm the primary role of R-loops in mediating SF3B1^{K700E}-mediated DDR.

To further characterize mechanisms of growth arrest in these cells, we performed flow cytometry for DNA content on isogenic cells, at 4 and 8 days after induction of ^{K700E} locus expression. As shown in Fig4D (and FigS4K), SF3B1^{K700E} and SF3B1^{WT} cell cycle distributions were similar on day 4, but by day 8, a significant fraction of SF3B1^{K700E} cells were noted to accumulate in S-phase compared to SF3B1^{WT}. Importantly this reduction in G_1 and increase in S-phase were abrogated by RNaseH1 overexpression. Co-staining for γH2AX, cleaved caspase 3, and 5-ethynyl-2′-deoxyuridine (EdU) also confirmed that DDR and apoptosis was limited to S-phase arrested cells (FigS4L). S-phase arrest in the absence of toxic or metabolic perturbation suggests replication stress wherein unstable replication forks cannot be resolved[59 60]. Since R-loops arise at sites of transcription[58 61], and splicing is co-transcriptional, we reasoned that SF3B1^{K700E} spliceosomes are linked to TRCs and DDR through transcriptional dysregulation. To confirm that SF3B1^{K700E} induced DDR requires the presence of active transcription bubbles, we used a double thymidine block-induced arrest to synchronize SF3B1^{K700E} cells at the G1/S checkpoint. Cells were subsequently released into the S-phase either in the presence or absence of triptolide, a RNAPII inhibitor[62]. DNA damage signaling was assayed at the single cell level, using γH2AX IF and TRCs by PLA, at 6 hours post-release (FigS4M). The effectiveness of triptolide in inhibiting transcription was confirmed using through 5-ethynyl-2′-uridine (EU) incorporation (FigS4M). Cells pretreated with triptolide showed abrogation of DDR and TRC formation, suggesting that DNA damage caused by SF3B1^{K700E} expression indeed requires the presence of active transcription bubbles. We finally analyzed whether increased R-loops in SF3B1^{K700E}impair DNA replication using two approaches. First, DNA replication fork analysis of combed DNA fibers (using sequential double pulse labeling with two thymidine

analogues (IdU and CldU))[63], showed higher de novo origins of replication (FigS4N), reduced RF velocity (Fig4E) and reduced DNA fiber track length (FigS4N) in SF3B1^{K700E} versus the SF3B1^{WT}. These effects were partially reversed by RNAseH1-OE. Second, incorporation of nucleoside analogue 5EdU (a measure of newly synthesized DNA)[64] was determined to be lower in SF3B1^{K700E} cells (Fig4F). This decrease was also partially normalized by RNaseH1-OE. Together, these results indicate that SF3B1^{K700E}- associated R-loops impair RF progression and induce replication stress.

R-loop accumulate primarily in gene bodies upon SF3B1^{K700E} expression

Slowed transcription elongation rate leads to gene-body R-loop accumulation and genome instability[65 66]. These RNAPII arrest-associated R-loop levels are further modulated by the orientation of their accompanying TRCs[67], occurring between the RNA and DNA polymerase machineries. We posited that the increased R-loops may be predominantly localized to gene bodies. To map R-loop distribution genome-wide, we immunoprecipitated and sequenced chromatin with S9.6 antibody (DRIP-seq)[68](FigS4O, P). DRIP-seq was performed at 4 days of ^{K700E} locus induction when there were no differences in cell-cycle distribution (Fig4D). DRIP-seq data revealed 4576 peaks more enriched in SF3B1^{K700E}, a ~350% increase compared to SF3B1^{WT} (FigS4Q-R). Gene-metaplot distribution analysis of R-loops showed progressively increasing signal intensity in gene bodies peaking at termination sites in SF3B1^{K700E} (Fig4G). Relative enrichment of R-loops at transcription termination sites (TTSs) is a feature of replication stress arising from HO-TRCs[67 69] , corroborating the increased TRCs and impaired RF progression. Since replisome movement usually orients in the same direction as transcription, HO-TRCs are particularly enriched in TTSs of gene neighbor pairs with convergently oriented gene transcription (convergent gene pairs) (FigS4S). Such convergent gene pairs are thus distinct from gene pairs which may be proximal but transcribe in the same or divergent directions (non-convergent gene pairs, FigS4S). We classified genes as convergent or non-convergent groups based on orientation of transcription between neighboring genes and constructed separate DRIP-seq metagene plots. As shown in Fig4H, gene-pairs with convergent transcription had markedly increased enrichment of R-loop signals at TTSs in SF3B1^{K700E} implying the occurrence of HO-TRCs further augmenting the R-loop levels[67] at these HO-collision hotspots. By contrast, non-convergent genes R-loop enrichment at TTS enrichment in SF3B1^{K700E} was much less pronounced (FigS4T). Furthermore, differential DRIP-signal increase at the TTS regions of converging gene pairs was dependent on distance between TTS (FigS4U). We next intersected DRIP-seq peaks with previously reported genome-wide RF directionality data (Okazaki fragment sequencing or OK-seq) from normal K562 cells[70] (FigS4V). SF3B1^{K700E}-specific genic R-loop peaks were checked for enrichment for RFs) – either HO or co-directional (CD). The absolute number of SF3B1^{K700E} DRIP-seq peaks (Fig4I) and mean DRIP-seq signal (Fig4J) were both higher in HO compared to CD conflict zones. The R-loop associated ssDNA accumulation at these stalled forks likely serve as the substrates for robust ATR-Chk1 DDR pathway activation (FigS4W)[46 49] in SF mutations). Taken together, our data show increased gene body R-loop

accumulation consistent with the gene-body elongation defect due to SF3B1^{K700E}, with R-loops particularly enriched at convergent gene-pair terminal sites correlating with HO conflict zones during DNA replication.

SF3B1^{K700E} triggers changes to histone marks and chromatin accessibility

Given the intricate functional link between R-loops, chromatin organization and transcription[71 72] and widespread epigenetic dysregulation observed even in SF mutant myeloid disorders[73 74], we determined changes to chromatin states effected by SF3B1^{K700E}. We used CUT&RUN sequencing[75] to map well defined histone marks associated with functional genomic states: H3K4me3 (promoter transcriptional activity), H3K27ac (active promoters and enhancers), H3K27me3 (repressive chromatin), and H3K4me1 (distal regulatory elements including poised enhancers) (FigS5A). Among these, H3K4me3 was prominently reduced at promoters in SF3B1^{K700E}, as evident by metaplots and heatmaps (Fig5A-B) as well as in descriptive examples of individual genes (Fig5C). H3K4me3 marks are now known to be dependent on RNAPII transcriptional activity (specifically p-p release) and efficient splicing at the first exon-intron boundary[76]. Reduced p-p H3K4me3 marks in SF3B1^{K700E} (FigS5B) may therefore arise from defective transcription or splicing at this region. Consistently, changes in H3K4me3 signal closely tracked corresponding changes in RNAPII ChIP-seq densities upon SF3B1^{K700E} expression (Fig5E), underscoring the role of RNAPII density in H3K4me3 distribution[77 78]. Among promoter regions, bivalent or "poised" promoters (enriched for both the active H3K4me3 and repressive H3K27me3 marks) showed decreased H3K4me3 and correspondingly increased H3K27me3 marks in SF3B1^{K700E} (Fig5D). Such a change favors a closed chromatin configuration in these regions[79]. Of note, H3K27ac promoter signal intensities did not differ between SF3B1^{WT} and SF3B1^{WT}(FigS5C).

To determine changes to chromatin accessibility and nucleosome organization, we performed ATAC-seq. An overall reduction in chromatin accessibility was evident in SF3B1^{K700E} compared to SF3B1^{WT} (5538 vs 126 significantly downregulated peaks, respectively) (Fig5F). Of the ATAC peaks downregulated in SF3B1^{K700E}, 44% (8.5% of all accessible promoters) were at promoters and 13% (1.1% of all accessible putative enhancers) at enhancers. R-loops may increase or decrease chromatin accessibility, depending on the context[80]. Given the preponderance of excess R-loops at gene bodies, we posited that changes to chromatin accessibility and R-loops may be localized to similar regions. Colocalization[81] of downregulated ATAC-seq peaks and differential DRIP-seq peaks in SF3B1^{K700E} showed enrichment of upregulated promoter and gene body DRIP-seq at closed chromatin regions (Fig5G, FigS5D). This suggests that excess R-loops are closely related to reduced chromatin accessibility in SF3B1^{K700E}. We next implemented NucleoATAC[82], an algorithm to infer nucleosome occupancy from ATAC-seq datasets to determine nucleosome occupancy at TSS. TSS of actively transcribing genes are devoid of nucleosomes; thus, increased nucleosome occupancy denotes transcriptional repression[83 84]. As shown in Fig5H, SF3B1^{K700E} had increased nucleosome occupancy at TSS, further

confirming transcriptional repression accompanying RNAPII elongation defects. GRO-seq RNAPII density at promoter and enhancers is a surrogate measure for their ability to activate transcription of their target genes[85]. We found loss of transcribing RNAPII activity extending from pausing sites (highest GRO-seq signal within the +20 to +300 bp window downstream of TSS) to downstream regions for the sense transcription, at both promoters (upstream region of TSS) and putative enhancers (distal H3K27ac peaks) (Fig5I, FigS5E-F). Taken together, our data show changes to histone marks and chromatin accessibility consistent with repressed RNAPII transcription.

We next applied a Hidden Markov Modeling (ChromHMM)[86], defining a set of chromatin states based on the histone modifications, DRIP-seq, and ATAC-seq, to understand how chromatin landscape alterations correlate with RNA-seq gene expression changes in SF3B1^{K700E} cells. A final model containing 14 states, annotated by intersecting each with known genome annotation features (FigS5G-H), was adopted for downstream analysis (Fig5J). Within each of these groupings, enrichment of specific genomic structures was as expected (FigS7G-H), supporting the biologic relevance of the model and assigned state annotations. Calculation of coverage changes in the genome-wide occupancies for each state, from SF3B1^{K700E} to WT, are shown in FigS5I. To specifically understand relationships between chromatin states and mRNA-seq gene expression, we calculated the relative enrichment of all possible chromatin state transitions at the promoters (+0.5 kb and −0.5 kb from TSS) of genes that were differentially expressed with an expectation that genes downregulated in SF3B1^{K700E} cells would show a switch from the most 'active' promoter state (state 13) to 'less active' states (states 10 - 12) on their promoters and vice-versa. As highlighted in Fig5K, (see also supplemental, Sourcedata_ChromHMM file for details), we observed a reciprocal pattern of chromatin state transition enrichments consistent with 'active' to 'less active' and 'less active' to 'active' in the SF3B1^{K700E} downregulated and upregulated genes, respectively. These data indicate the combinatorial effects of transcription kinetics, R-loop distribution and changes to chromatin landscape on the steady state transcriptome of SF3B1^{K700E} cells.

Components of the Histone Deacetylase/H3K4me pathways modulate sensitivity to SF3B1^{K700E}

Based on the changes to H3K4me3 marks and chromatin accessibility, as above, we speculated if modulating epigenetic regulators may improve RNAPII elongation defects and rescue cells from downstream effects including growth arrest. We hence performed a forward genetic screen with a short hairpin RNA (shRNA) library targeting epigenetic regulator and chromatin modifier protein classes (~2500 shRNAs targeting 357 genes, Fig6A). SF3B1^{K700E} and SF3B1^{WT} cells were first stably transduced with the shRNA library and selected with puromycin for 7 days. SF3B1^{K700E} was then induced by addition of doxycycline and cells sampled at 8-day intervals to determine changes to abundance of shRNA (FigS6A). We focused our analyses on those shRNAs that increased in abundance selectively in the SF3B1^{K700E} ("survival candidates") since knockdown of the candidate genes would improve cell survival and growth (Fig6B; TableS1). The top 20 survival candidates were noted to be highly enriched in the histone deacetylase (HDAC) pathway by GO

analysis (FigS6B). Analysis for protein-protein interactions using the StringDB package showed a similar enrichment of Sin3/HDAC complex (Fig6C). Specifically, the top hits (HDAC2, ING2, and PHF21A (or BHC80)) are part of the extended LSD1-Sin3-HDAC transcriptional corepressor complex which negatively modulates H3K4me activity[87 88 89 90 91 92 93 94]. Additionally, WDR5, a key writer component of H3K4me3 methyltransferase complex[95], was depleted in the screen (i.e., WDR5 loss led to accelerated cell death in SF3B1^{K700E} cells) (TableS2).

Secondary validation experiments with two shRNA each for HDAC2, ING2 and PHF21A confirmed improved cell survival (Fig6D), upon knockdown of these genes (FigS8C). Next, we determined that the improved cell survival was associated with a reduction in R-loops, DDR and TRCs (using nuclear S9.6, tail moments, γH2AX and PLA nuclear signals, respectively (FigS6D-F, Fig6E). This also corresponded to restoration of cell cycle distribution of SF3B1^{K700E} (FigS6G), with improvement in S-phase replication dynamics (FigS6H-I). We then asked if improved S-phase replication dynamics was accompanied by correction of RNAPII transcription elongation kinetics. The RNAPII Ser2P ChIP profiles showed reduction in the gene body densities with loss of HDAC2, ING2 or PHF21A in SF3B1^{K700E} cells (Fig6F), suggesting improved RNAPII transcription kinetics as the likely mechanism underlying the reduced TRCs in these cells. Finally, we asked whether improvement in RNAPII transcription kinetics in these cells would also be reflected in the normalization of global distribution of promoter H3K4me3 marks. As shown in Fig6G, there was a significant increase in H3K4 trimethylation observed in the shRNA- protein knockdown SF3B1^{K700E} cells, confirming that loss of these modulators affected H3K4me3 mark deposition in the promoter region either directly or through improvement in RNAPII elongation kinetics. Taken together, our results suggest that transcription changes induced by SF3B1^{K700E} can be mitigated by modifying chromatin states.

To address how RNAPII occupancy relates to selection of cryptic 3'SSs in SF3B1^{K700E}, specifically the role of transcription speed in the selection of cryptic 3'SSs, we selected cryptic 3'SSs (from both our K562 isogenic cells and analysis of patient samples[15]) and determined RNAPII distribution (from ChIP-seq data) in introns containing such cryptic 3'SSs. The introns (or genes containing such introns) were no different from unselected introns (or unselected genes) in their SF3B1^{K700E} RNAPII distribution changes (FigS6J-K). To further understand impact of transcription on SF3B1^{K700E} -induced cryptic 3'SSs usage, we performed RT-PCR on select introns containing cryptic 3'SSs and found that SF3B1^{K700E} cryptic 3'SS usage did not decrease upon improvement of transcription kinetics (such as through HTATSF1-OE, shHDAC2 or shING2) (Fig6H). These findings suggest that altered transcription and cryptic 3'SS selection in SF3B1^{K700E} is likely driven by independent mechanisms.

WDR5 inhibition reduces colony formation in SF3B1-mutant MDS

We analyzed CD34+ cells isolated from up to seven SF3B1-mutant MDS patient samples (and 5 CD34+ controls from normal age-matched donors) to confirm key findings from the isogenic cell line system (Fig7A

& FigS7A). MDS patient-derived CD34+ cells showed significantly higher TRCs (FigS7B) and DDR (FigS7C) by gamma-H2AX and PLA microscopy, respectively. Given that ChIP-seq is difficult to perform with low cell numbers, we optimized CUT&RUN for Ser2P RNAPII to map RNAPII distribution. Metagene analyses of Ser2P RNAPII occupancies revealed a redistribution of RNAPII with reduced promoter/TSS density and increased gene body levels (Fig7B-C) in MDS samples, similar to isogenic cells. Having confirmed RNAPII redistribution, we then determined if similar changes to histone profile and chromatin accessibility also exist in MDS samples. CUT&RUN for H3K4me3 demonstrated a significant reduction in promoter H3K4me3 densities in SF3B1-mutant MDS (Fig7D-E). Chromatin accessibility was additionally assayed in the same cells with ATAC-seq. Nucleosome mapping on the ATAC-seq datasets also showed changes in-line with isogenic cell systems (reduction in nucleosome occupancy at the −1 and +1 positions accompanied by higher occupancy at the nucleosome-depleted region) (Fig7F). Our patient sample data thus confirms changes to transcription kinetics and chromatin organization in primary patient samples as predicted by isogenic SF3B1^{K700E} cells.

We then sought to define the functional relevance of reduced H3K4me3 promoter marks and H3K4me functional pathway on cell growth kinetics in SF3B1^{K700E}. Based on the shRNA screen results (Fig6), we determined how inhibition of WDR5 affects *in vitro* colony growth of SF3B1-mutant MDS using OICR-9429, a small molecule inhibitor of WDR5 with anticancer efficacy against non-MLL-rearranged leukemia and certain solid tumor cancers[96 97 98]. As shown in Fig7G, treatment with OICR-9429 at the previously established IC_{50} of 5uM[99 97] had a modest effect on colony growth in WT healthy volunteer replicates. By contrast, OICR-9429 caused a significant decrease in colony numbers of most SF3B1-mutant MDS colonies. We also ascertained the effect of shRNA-mediated knock-down of ING2 and HDAC2, two of the top targets in the shRNA screen on colony growth. As in the isogenic cell line culture, loss of either HDAC2 or ING2 knockdowns significantly promoted colony formation of SF3B1-mutant MDS cultures; the effects were more pronounced upon ING2 knockdown (Fig7H, FigS7D). Bulk DNA sequencing of the colonies confirmed mutant colony growth (FigS7G). Flow cytometry evaluation of these colonies revealed an increase in proportion of both mature myeloid (CD11b) or erythroid (CD71) markers in colonies compared to shRNA-transfected controls from the same donors (FigS7E-F). Thus, knocking down ING2 or HDAC2 promotes lineage commitment and differentiation. Together, our results point towards the potential to therapeutically target specific epigenetic pathways in SF3B1-mutant MDS.

We finally utilized a previously established conditional murine model of Sf3b1^{K700E}[100] of to independently confirm its functional interplay with Sin3/HDAC pathway in primary cells. The Sf3b1^{K700E} mice were crossed with Tamoxifen-Cre mice, and double mutant mice (*R26*-CreER$^{T2\ KI/+}$ *Sf3b1*$^{+/K700E}$) were induced by intraperitoneal injections with tamoxifen. LSK progenitor population (cKit+Sca-1+Lin-) were isolated from the bone marrow of the double mutant mice (along with wild-type age-matched controls (*R26*-CreER$^{T2\ KI/+}$

Sf3b1$^{+/+}$)) (Fig7I, FigS7H) and transduced with lentiviruses expressing shRNAs targeting Hdac2, Ing2, and Wdr5, or scrambled control. The ability of cells to form colonies was assayed by standard methyl-cellulose CFU assays at 14 days of cultures. Consistent with observations from human patient sample data, we observed a selective increase in colony growth with loss of Hdac2 and Ing2 (Fig7J,FigS7I-J). Conversely, Wdr5 knock-down inhibited colony growth selectively in the mutant cells (Fig7J, FigS7I). We also performed ATAC-seq on LSK cells to determine changes to nucleosome occupancy. As shown in Fig7K, NucleoATAC analysis showed a similar pattern of nucleosome occupancy in mutant murine cells as was observed for both patient samples and isogenic cells. These findings in the murine model provides additional credence to WDR5/H3K4me3 pathway, representing a targetable functional vulnerability in SF3B1-mutant diseases.

DISCUSSION

The molecular mechanisms through which mutant SFs serve as oncoproteins have been extensively investigated. Most of these studies have focused on changes in the isoform ratio of well-defined tumor suppressors or oncogenes through AS[101 102]. We independently analyzed 172 MDS/AML patient samples with SF mutations and 45 wild-type controls, curated from three different studies[33 35 34] to determine patterns of AS associated with mutational subtypes (FigS10). Consistent with previous reports, we found evidence of widespread AS events that clustered only loosely along the mutational subtypes. However, there were few AS events common across the different SF mutations that could explain the common MDS phenotype. This prompted the current study to identify AS-independent mechanisms. One such is the increased presence of R-loops, its downstream activation of Chk1-ATR pathway, and DDR observed across the various SF-mutation subtypes[46 49]. Using a genome-wide strategy based on the IP of catalytically inactive RNaseH (R-ChIP), Chen et al. showed that R-loops increased upon inducible overexpression of mutant SRSF2 or U2AF1, largely in promoter regions[46]. For SRSF2, a reduced ability of the mutant protein to extract pTEFb from the inhibitory complex 7SK during transcription elongation was proposed to underlie the impaired promoter pause release of RNAPII and excess promoter R-loop accumulation. This interaction being specific to SRSF2 is not applicable to other SF mutations, namely SF3B1 or U2AF1.

Given that transcription and splicing are closely coordinated, we used an inducible isogenic model system to determine changes in transcription associated with SF3B1 and U2AF1 mutations through a variety of approaches. ChIP-seq of RNAPII (both nonspecific RNAPII and elongation-specific RNAPII) confirmed an elongation defect that increased along the length of the gene body until TTS, with a decrease downstream of TTS. We then used two transcriptomic approaches (TT-TL-seq and GRO-seq) to determine changes in nascent transcription, both which confirmed the ChIP-seq results. RNAPII elongation defects were most pronounced in genes with more introns and greater lengths. Additionally, the RNAPII elongation defects were distinct from those arising due to generalized inhibition of splicing. There was no evidence of an increased readthrough in any of our nascent transcriptome assays that would suggest an accompanying termination defect. Importantly, RNAPII accumulation was predominantly along introns compared to exons, as expected with perturbation of U2 snRNP-dependent RNAPII elongation. Consistent with the RNAPII transcription elongation defect, we found increased R-loops along gene bodies, TRCs and DNA damage response. Isogenic U2AF1^{S34F} mutant cells also recapitulated key aspects of RNAPII elongation defects, suggesting that dysregulation of transcription kinetics may be shared across SFs and may underlie their mutual exclusivity.

Our findings that place RNAPII elongation defects that linked SF-mutations to increased R-loops and replication stress prompted us to ask the follow up questions:

Splicing is co-transcriptional and SF mutations are associated with distinct patterns of AS; hence how are splicing and RNAPII elongation defects related? SF3B1 mutations are associated with selection of novel or "cryptic" 3'SS, and also associated with a decrease in retained introns[30 14] (as determined by RNA-seq). Hence, our analysis focused on how RNAPII occupancy is related to the selection of cryptic 3'SSs and retained introns. We selected cryptic 3'SS (from both isogenic cells and patient samples) and determined the RNAPII distribution (from ChIP-seq data) of introns containing such cryptic 3'SSs. These introns did not differ from unselected introns in their RNAPII distribution. Next, we studied CoSE, a metric to determine the efficiency of intron splicing in nascent RNA. If reduction of RI is related to improved splicing efficiency, an increased CoSE is expected. SF3B1 mutations were associated with a reduction in CoSE, suggesting that the reduced RI does not arise from the improved efficiency of co-transcriptional splicing. Taken together, our data suggest that changes in alternative splicing and RNAPII transcription are independent. While the use of cryptic 3'SS appears to be related to use of non-canonical BP, the basis of reduced RI is unclear and likely results from improved processing of nuclear-detained RI, which requires further study.

Given the interplay between transcription and chromatin states, how do RNAPII elongation defects affect chromatin organization and vice versa? Interdependence of transcription on histone marks, DNA methylation and chromatin organization has been well described[72 103 104]. Furthermore, we found that SF3B1^{K700E} associated transcription changes were accompanied by changes in histone marks (predominantly H3K4me3) as well as chromatin accessibility. We then used an shRNA library against epigenetic regulators: the dramatic reduction in cell viability upon induction of SF3B1^{K700E} expression provided a robust survival endpoint for this assay. The results of this screen strongly implicate the Sin3/HDAC pathway in regulating RNAPII elongation and its downstream effects. These results also hold true in primary patient samples and a mouse model of Sf3b1^{K700E}. Our results linking transcription to chromatin organization are significant for many reasons. First, MDS is broadly susceptible to epigenetic therapies regardless of the type of driver mutation. Our results suggest that even with SF-mutant disease without genetic alterations to epigenetic factors, mechanisms may converge upon epigenetic mechanisms. Second, these results may explain why SF mutations are largely exclusive to some epigenetic mutations (such as those in the PRC2 complex)[105,106]. One possibility is that PRC2 and SF mutations may have similar effects on chromatin organization rendering dual mutant cells unviable. The role of 3D genome organization in clonal myeloid disorders has recently been explored[107], and our results underscore the need for mechanistic studies that link oncogenic drivers with altered chromatic organization.

Are the changes to transcription and chromatin demonstrable in human patient samples and transgenic mouse models, and can the modulation of RNAPII elongation defects be exploited therapeutically? Given the limited cell numbers available from patient samples and primary mouse cells, we chose to replicate key results from the isogenic cells with assays that allowed for low cell numbers (CUT&RUN, ATAC-seq, and

colony growth assays). Critically, both models confirmed key findings, including survival differences, when epigenetic factors were knocked down using shRNA directed at the Sin3/HDAC complex or a small molecule inhibitor (against WDR5). WDR5 is a key "writer" in the COMPASS H3K4me3 methyltransferase complex, which was depleted in the shRNA rescue screen, suggesting a lethality due to its loss in SF3B1^{K700E} cells. Indeed, chemical inhibition of WDR5 resulted in profound inhibition of colony growth in patient samples as well as in mutant mouse cells without affecting wild-type controls. WDR5 inhibitors are currently explored for their anti-tumor effects broadly[108 96 97 98]. Our results suggest they may be particularly effective in situations where RNAPII transcription elongation is disrupted.

What are the molecular mechanisms through which mutant SFs affect RNAPII elongation? Recent cryo-EM studies have shed light on the organization of early spliceosome complexes[36 37]. Accordingly, a critical step in the assembly of the pre- spliceosome A complex is the eviction of HTATSF1 (human ortholog of yeast Cus2) bound to SF3B1 by DDX46 (human ortholog of yeast Prp5), which allows the U2AF complex to bind to SF3B1[36 37]. Mutations in Prp5 or U2snRNA induce an RNAPII elongation defect in introns[109], suggesting U2snRNP assembly-dependence of transcription elongation activation. Separate studies have also shown the direct interaction of RNAPII and U1 snRNP during early spliceosome assembly; RNAPII is released only upon the formation of the pre-B complex on its surface (addition of U4/U6.U5 snRNP to the A-complex)[2]. Investigating this further, we found that HTATSF1 binds less avidly to SF3B1^{K700E} than SF3B1^{WT}. *In vitro* splicing analysis showed reduced progression of SF3B1^{K700E} spliceosomes from the E to the A complex. Importantly, disruption of early spliceosome assembly by knocking down either HTATSF1 or DDX46 results in a similar pattern of RNAPII elongation defect as that observed with SF3B1^{K700E}. Overexpression of HTATSF1 also resulted in partial rescue of growth arrest of SF3B1^{K700E} isogenic cells. Taken together, we propose a model in which impaired interaction of SF3B1^{K700E} with HTATSF1[110 111] prevents DDX46-induced change to the confirmation of SF3B1 to evict HTATSF1. This prevents SF3B1-U2AF interaction critical for the E to A complex transition impairing the release of RNAPII from its bound configuration to U1 snRNP (Fig8E).

In summary, our results define how RNAPII elongation is disrupted in response to oncogenic mutations in SF3B1 and U2AF1. Modulation of epigenetic regulators improves RNAPII transcription kinetics and point to potential therapeutic targets. An important inference from our model is that the RNAPII elongation defects arise from the defective assembly of early spliceosomes. Interestingly, most SFs mutated in disease, including SF3B1, U2AF1, SRSF2, U1 snRNA, have established roles in early spliceosome assembly[112], suggesting a possible shared defective RNAPII transcriptional response to these mutations. Exploring RNAPII kinetics and early spliceosome assembly in these additional mutational contexts will be informative for defining potential therapeutic targets similar to the Sin3/HDAC pathway in *SF3B1* mutations.

Limitations of the study:

This study primarily utilized an ENCODE tier 1 cell line, K562, for genome-wide studies and screens. We acknowledge that specific genes or pathways relevant to oncogenesis cannot be inferred from transformed cell lines. However, the fundamental aspects of transcriptional regulation are likely preserved across cell systems. Critical aspects of our work were confirmed in human samples and mouse models, an approach used in many previous studies. Although most of the sequencing data were collected at the earliest optimal time point of 4 days following doxycycline induction, we cannot entirely rule out indirect or secondary effects. LRS studies are currently limited by the low total read depth on the PacBio platform. Finally, additional confirmations for the working model we propose may require additional comprehensive studies utilizing single molecule microscopy or molecular structure.

MATERIALS AND METHODS

Antibodies: Antibodies used in the study are listed in Table S8

Cell lines:

All parent cell lines were obtained from ATCC unless otherwise stated, authenticated by ATCC STR profiling, and tested for mycoplasma contamination at regular intervals using Mycoplasma PCR detection kit (Boca Scientific). K562 were grown in RPMI1640 medium supplemented by 10% fetal bovine serum (FBS) and 1% penicillin-streptomycin. HEK293T cells were grown in DMEM supplemented by 10% FBS and 1% penicillin-streptomycin. *Drosophila melanogaster* S2 (Schneider-2) cells (used for TT-TL seq and ChIP-seq spike-ins) were cultured in Schneider's modified *Drosophila* medium (Lonza) supplemented with 10% FBS at 27°C without CO_2. Doxycycline induction was performed using Doxycycline (Sigma) at 1 ug/ml concentration respectively. Puromycin, blasticidin, hygromycin, and neomycin antibiotic resistance selections and growth were performed at 1 ug/mL, 10 ug/mL, 250 ug/mL, and 200 ug/mL concentrations, respectively. Isogenic $SF3B1^{K700(E/K)}$ and $U2AF1^{S34(F/S)}$ endogenous knock-in K562 cells were generated using a combinatory AAV intron trap-CRISPR approach (sgRNA/CRISPR mediated DSB and AAV-intron trap-vector donor template) as previously described[21].

siRNA and shRNA mediated knock-down:

For single siRNA knockdown K562, cells were electroporated with 50 nM siRNA DharmaFECT 1 (Dharmacon) per manufacturer's instructions (Nucleofector V kit, T-106 program). Pooled siRNAs (ON-TARGET SMARTpool siRNAs from Dharmacon) were used for knockdown of UAP56, SETX, FANCD2, DDX46, and HTATSF1 were used to achieve protein depletion. ON-TARGETplus Non-targeting Control Pool was used as the control (siC). A detailed list of siRNAs is in Table S3. Gene expression changes were assessed by qRT-PCR using gene-specific primers (Table S7).

Lentiviral gene-specific shRNAs were prepared in pLKO.1-puro backbone (Addgene # 8453) (identified from verified sequences at https://www.sigmaaldrich.com/US/en/product/sigma/shrna; TableS4). Lentiviral doxycycline-inducible shRNAs (targeting DDX46 and HTATSF1) were prepared in the Tet-pLKO-puro backbone (Addgene # 21915) by cloning the gene-targeting annealed oligos also identified from the above url into the AgeI - EcoR1 sites (details in TableS5). Lentivirus supernatants were prepared by co-transfecting the Tet-pLKO-puro plasmid along with helper plasmids pVSV-G (Addgene #138479) and p8.92 (Addgene #132929) into 293T cells as previously described. Supernatants were used as is or concentrated by PEG-8000 if required. Cells were spinfected (1800 RPM for 45 minutes at room temperature) in growth media supplemented with Polybrene (Yeasen), incubated at 37 C for 48 hours before selecting in 1 μg/ml puromycin for 5-7 days.

Inducible lentiviral overexpression:

Doxycycline-inducible gene overexpression vectors for DDX46 and hTATSF1 were prepared by cloning C-terminally HA-tagged DDX46 (Genscript, OHu29752) and hTATSF1 (amplified from K562 cDNA prepared using oligo(dT) primers) ORFs respectively into pLIX_403 (Addgene # 41395). Viral preps and selection were done as in previous section. Gene expression changes were assessed by qRT-PCR using transcript-specific primers (TableS7).

Quantification of mRNA:

RNA purification was performed using the RNeasy Mini kit (QIAGEN) according to the manufacturer's conditions. cDNA synthesis was achieved using the M-MuLV Reverse Transcription kit (NEB), according to manufacturer's guidelines. Finally, qPCR was performed on a 7500 FAST Real-Time PCR system (Thermo Fisher Scientific) and messenger RNA expression values calculated using the $\Delta\Delta C$t method and *Beta-Actin* housekeeping gene as a control. A detailed list with primers used in the present study is provided in Supplementary TableS7.

Affinity purification of SF3B1-associated proteins:

For FLAG-tagged expression and purification in HEK293T cells, 1.5 million cells per plate were seeded in 10-cm plates. Next day, 5 ug of expression plasmid (expressing either Flag-tagged $SF3B1^{WT}$ or FLAG-$SF3B1^{K700E}$) were transfected using TransIT-X2 (MirusBio) and harvested after 48 hours. One-third was used for total nuclear fraction and two thirds for chromatin fraction by affinity purification. Protease inhibitor cocktail (Millipore Sigma, 11836170001) and phenylmethylsulfonyl fluoride (PMSF) were added to all the buffers to prevent protein degradation. For total nuclear fraction, cells were first lysed in cell lysis buffer (lysed in 250 µL cytoplasmic lysis buffer (10 mM Tris-HCl pH 7.5, 0.05% NP40, 150 mM NaCl) by gently resuspending then incubating on ice for 5 minutes. Lysate was then layered on top of a 500 µL cushion of 24% sucrose in cytoplasmic lysis buffer and spun at 2,000 rpm for 10 min at 4°C. The supernatant (cytoplasm fraction) was removed, and the pellet (nuclei) were rinsed once with 500 µL PBS/1 mM EDTA. Nuclei were resuspended in a 'NET' resuspension buffer (40mM Tris pH7.5, 150mM NaCl, 0.05% NP-40, 5mM MgCl2). Chromatin-bound SF3B1-protein fraction was isolated using a previously published protocol[113], with minor modifications. Briefly, the cell pellet was first lysed with 5 volumes of ice-cold E1 buffer for 1 volume of cell pellet and then centrifuged at 1,100 × *g* at 4 °C for 2 min. The supernatant (cytoplasmic fraction) was discarded, and the pellet resuspended with the same volume of E1 buffer, for a total of 2 washes, followed by incubation in E1 buffer for 10 min on ice and centrifugation at 1,100 × *g* at 4 °C for 2 min. The pellet was then resuspended in 2 volumes of ice-cold E2 buffer, centrifuged at 1,100 × *g* at 4 °C for 2 min, and the supernatant

(nucleoplasmic fraction) discarded. The pellet was again resuspended with the same volume of E2 buffer used for the previous step, for a total of 2 washes, followed by incubation in E2 buffer for 10 min on ice and centrifugation at 1,100 × *g* at 4 °C for 2 min. The chromatin pellet was resuspended in a resuspension buffer (40mM Tris pH7.5, 150mM NaCl, 0.05% NP-40, 5mM MgCl2). Both the nuclear and chromatin pellet fractions were sonicated 3 times (5 seconds on, 20 seconds off; 30% amplitude; Sonic Dismembranator Model 500 (Fisher Scientific)) on ice. For co-immunoprecipitation of FLAG-tagged SF3B1, nuclear extracts were incubated with ANTI-FLAG® M2 affinity gel (Sigma-Aldrich) for 8 hours at 4°C with gentle shaking, beads washed five times with pre-chilled NET resuspension buffer and eluted under native conditions by competition using 3x FLAG peptide (Millipore Sigma, F4799). Samples were boiled in the Laemmli SDS sample buffer (Thermofisher) at 95 C for 5 minutes.

Isolation and analyses of chromatin associated proteins:

All steps were performed on ice, and all buffers contained PMSF and 1x Roche protease inhibitor mix. Briefly, 40 million cells per sample were rinsed once with PBS/1 mM EDTA, then lysed in 250 µL cytoplasmic lysis buffer (10 mM Tris-HCl pH 7.5, 0.05% NP40, 150 mM NaCl) by gently resuspension, and incubated on ice for 5 minutes. Lysate was then layered on top of a 500 µL cushion of 24% sucrose in cytoplasmic lysis buffer and spun at 2,000 rpm for 10 min at 4°C. The supernatant (cytoplasm fraction) was removed, and the pellet (nuclei) were rinsed once with 500 µL PBS/1 mM EDTA. Nuclei were resuspended in 100 µL nuclear resuspension buffer (20 mM Tris-HCl pH 8.0, 75 mM NaCl, 0.5 mM EDTA, 0.85 mM DTT, 50% glycerol) by gentle flicking, then lysed by the addition of 100 µL nuclear lysis buffer (20 mM HEPES pH 7.5, 1 mM DTT, 7.5 mM $MgCl_2$, 0.2 mM EDTA, 0.3 M NaCl, 1 M Urea, 1% NP-40), vortexed for twice for 2 seconds each, and then incubated on ice for 3 min. Chromatin was pelleted by spinning at 14,000 rpm for 2 min at 4°C. The supernatant (nucleoplasm fraction) was removed, and the chromatin was rinsed once with PBS/1 mM EDTA. Chromatin was immediately dissolved in 100 ul of resuspension buffer (40mM Tris pH7.5, 150mM NaCl, 0.05% NP-40, 5mM MgCl2). Both nucleoplasmic and chromatin fractions were sonicated 3 times (5 seconds on, 20 seconds off; 30% amplitude; ultrasonic probe sonicator) on ice. Protein concentrations were determined using BCA assay and sample volumes adjusted. Samples were boiled in Laemmli SDS sample buffer (Thermofisher) at 95 C for 5 minutes.

In vitro splicing assay:

For SF3B1 overexpression in HEK293T cells, 1.5 million cells per plate were seeded in 10-cm plates for a total of 4 plates per condition. Next day, 5 ug of expression plasmid, per plate, (expressing either $SF3B1^{WT}$ or $SF3B1^{K700E}$) were transfected using TransIT-X2 (MirusBio) and cells were harvested after 48 hours. Small-

scale nuclear extracts (NEs) were prepared using a previously published protocol[114]. Protein concentrations in the dialyzed NEs were confirmed to be similar between the two conditions. The pSP72-AdML plasmid (Addgene#11242) was linearized using BamH1 enzyme digestion and the linearized template was purified by running in an agarose gel. [^{32}P]-labelled AdML pre-mRNA was synthesized in vitro using the T7 HiScribe Transcription Kit (NEB) using [α-^{32}P]UTP (Perkin Elmer, BLU007H250UC), and purified using ethanol precipitation. The NEs were mixed with either DMSO (both SF3B1^{WT} or SF3B1^{K700E}) or PlaB (Cayman Chemical, Cay16538-100) in DMSO and pre-incubated at 30°C for 10 minutes. Splicing reactions contained 15 ul of NEs diluted in dialysis buffer (20mM HEPES-KOH (pH8), 100mM KCl, 0.2mM EDTA, 20% glycerol) supplemented with 3mM $MgCl_2$, 1 mM ATP, 10 mM creatine phosphate, 2.6% PVA (added at the last), 1ul of Murine RNAase Inhibitor (40U/uL), and 10ng of [^{32}P]-labelled pre-mRNA in a total of 25 ul, per reaction. Splicing reactions were then incubated at 30 °C for the indicated time points. For the kinetic analysis of the spliceosome complexes, 5 ul of each reaction was mixed with 2.5ul of 10x Heparin loading dye (6.5mg/ml Heparin, 40%w/v glycerol, 0.5% bromophenol blue, 0.5% xylene cyanol) and incubated on ice, and the reactions were resolved on 1.5% agarose gel in 0.5x TBE by native gel electrophoresis for 3.5 hours at 70V. The agarose was subsequently fixed in a fixation solution (10% acetic acid, 10% methanol in water). Gels were dried in vacuum gel dryer at 80°C for up to 2 hours. The gels were scanned using a Typhoon phosphorimager (Molecular Dynamics).

Cell Synchronization and analysis:

K562 cells were synchronized as previously reported. Briefly, for G_2/M, cells were cultured in 40 ng/ml nocodazole (Sigma-Aldrich) for 20 h. For G_1/S arrest, double thymidine blockade was used (2mM Thymidine for 20 h, washed and recultured for 9 hours, and blocked a second time for 20 h). Cell cycle status was verified by FACS flow analysis.

Flow cytometry and cell cycle studies:

CytoPhase Violet (Biolegend) was used for DNA content evaluation by FACS by staining for 90 min per manufacturer's instructions. For immunophenotyping, cells washed in PBS/EDTA/BSA were typically incubated for 30minutes at 4 degrees and washed in same buffer 3 times. Antibodies used for FACS analysis and flow sorting are listed in TableS8. At least 20,000 events were recorded on a FACS LSR II HTS-1 flow cytometer (BD Biosciences) and analyzed using FlowJo software.

Cell growth kinetics:

To determine cell growth kinetics, equal number of cells (1 million per condition) were plated in 3 replicates. Cell counts were performed using a Nexcelom cellometer cell counter every XX days. Cell viability was independently determined by quantifying cellular ATP, using Cell Titre Glo 2.0 Cell Viability Assay (Promega) per manufacturer's instructions.

DNA fiber combing assay:

DNA combing was performed as previously described[63]. Briefly, exponentially growing K562 cells (3.0 x 10^5 per condition) were labeled with consecutive pulses of thymidine analogues (IdU and CldU) for 30-min. After labeling, cells were harvested, embedded in agarose and DNA was prepared then combed onto silanized coverslips using the FiberComb Molecular Combing System (Genomic Vision). Following combing, IdU and CldU were detected using mouse anti-BrdU (BD Biosciences 347580) and rat anti-BrdU (Abcam ab6326) antibodies, respectively. Sheep anti-mouse Cy3 (Sigma C2181) and goat anti-rat Alexa Fluor 488 (Invitrogen A11006) were used as secondary antibodies. Single-stranded DNA was counterstained with anti-ssDNA mouse antibody (DSHB University of Iowa), followed by anti-mouse BV480 (Jackson ImmunoResearch 115–685-166) as secondary antibody. DNA fiber images were acquired using an Olympus FV3000 confocal microscope. Track lengths were measured with ImageJ. To calculate replication fork velocity (kb/min), the following equation was used to convert fork length from mm to kb/min: length mm ´ 2/labeling time in min = fork speed kb/min (conversion factor of 2 kb/mm specific for DNA combing method). For quantification of fork velocity, a minimum of 150 fibers were analyzed per dataset.

Proximity Ligation assay:

PLA was performed as previously described using Duolink PLA Technology (Merck)[52]. Briefly, cytospin slides were pre-extracted, fixed, permeabilized and incubated with primary antibodies as described above for IF. Secondary antibody binding, ligation and amplification reactions were then performed according to the Duolink instructions. Duolink in situ PLA probe anti-rabbit PLUS (Merck) and Duolink in situ PLA probe anti-mouse MINUS (Merck), and Duolink Detection Reagents Red or Green (Merck) were used to perform the PLA reaction. Finally, nuclei were stained with DAPI and mounted in ProLong Gold AntiFade reagent (Cell Signaling Technology, 8961S). For PLA reactions requiring mouse FANCD2, mouse PCNA, rabbit PCNA, or rabbit RNAPII NTD antibodies, 1:200 dilution of these antibodies was used. Finally, images were acquired with a SP8 STED microscope equipped with a DFC390 camera (Leica) at 60× magnification and LAS AX image acquisition software (Leica). A FIJI image processing package[115] was used for image analysis and quantification. PLA foci numbers per cell were quantified for all conditions.

EdU and EU incorporation and detection

EdU (Catalog No. C10337) and EU (Catalog No. C10329) labeling were performed using Click-IT chemistry kits from Invitrogen. First, cells were cultured in complete medium supplemented with 10 μM EdU for 30 min. Then, samples were fixed, permeabilized and Click-iT reaction performed according to the manufacturer's guidelines. Finally, nuclei were stained with DAPI and mounted in ProLong Gold AntiFade reagent

(Invitrogen). Images were acquired with a Leica DM6000 microscope equipped with a DFC390 camera (Leica) at 63× magnification and LAS X image acquisition software (Leica). The FIJI image processing package was used for image analysis and quantification. The EdU entire population nuclear intensity and percentage of cells incorporating EdU were determined. EdU intensity was also determined only for those cells that incorporated EdU.

Alkaline single-cell electrophoresis:

Alkaline single-cell electrophoresis or comet assay was performed as previously described[116]. Comet slides were stained with SYBRsafe, and images were captured with a Leica DM6000 microscope equipped with a DFC390 camera (Leica). Analysis of comet assay images was performed using FIJI. Tail moment was measured for each cell in each condition.

Immunofluorescence (IF):

K562 cells were cytospun at 400 RPM x 3 minutes onto glass coverslips coated with Poly-L-lysine (Gibco) in a Thermoshandon Cytospin 3 centrifuge. S9.6 IF was performed essentially as previously described[51]. Briefly, cells were fixed with 100% ice-cold methanol x 15 minutes, blocked with phosphate-buffered saline–bovine serum albumin (PBS–BSA) 2% overnight at 4 °C and incubated with S9.6 (1:1,000) and anti-nucleolin (1:2,000) antibodies overnight at 4 °C. Then, coverslips were washed three times in PBS 1× and incubated with secondary antibodies (1:1,000) for 1 hr at room temperature (RT). Finally, cells were washed again, stained with DAPI (1:1000, Thermofisher) and mounted in ProLong Gold AntiFade reagent (Cell Signaling Technology, 8961S). IF images were acquired with a Leica DM6000 microscope equipped with a DFC390 camera at 63× magnification and LAS X image acquisition software (Leica). A FIJI (ImageJ) image processing package was used for IF analysis. The nuclear mean gray value for S9.6, after subtraction of nucleolar signal, was measured for each condition.

DNA-damage assessment by γH2AX immunostaining was performed mainly as previously described[117] with minor modifications. All cells were fixed in paraformaldehyde for 20 min, washed twice in PBS, permeabilized with PBS containing 0.1% Triton-X100, and blocked in PBS + 2% bovine serum albumin. Immunofluorescence was performed using anti-γH2AX (1:1000) primary antibody overnight at 4 °C. Coverslips were washed three times with PBS and incubated with Alexafluor secondary antibodies (1:1000) and counterstained with DAPI (1:1000, Thermofisher) for 1 h at room temperature. Coverslips were mounted on glass slides using ProLong Gold anti-Fade mount (Cell Signaling Technology, 8961S) and cured overnight. IF images were acquired with a Leica DM6000 microscope equipped with a DFC390 camera at 63× magnification and LAS AX image acquisition software (Leica). A FIJI (ImageJ) image processing package was used for IF analysis.

Dot-blot assay:

Cells were lysed in 0.5% SDS/TE, pH 8.0 containing Proteinase K overnight at 37 °C. Total chromatin was isolated with phenol/chloroform/isoamyl alcohol extraction followed by standard ethanol precipitation. The eluted chromatin was digested for one hour in RNaseIII (NEB, M0245), purified with MinElute PCR purification kit (Qiagen, 28004), and quantified using Nanodrop. A negative control treated for 4 hours at 37 °C with RNaseH1 (New England Biolabs) was included for each condition. 300 and 600 nanograms of total DNA was loaded in duplicate onto a Nylon membrane using dot blot apparatus, UV crosslinked wit at 0.12 Joules and stained with methylene blue to determine uniform loading. Membranes were then blocked with 5% skim milk in PBST (PBS; 0.1% Tween-20) for 1 hour and RNA:DNA hybrids were detected by immunoblotting with S9.6 antibody.

ChIP-seq:

ChIP-seq was performed as previously described[118] with 1-5 × 10^7 cells from two biological replicates. Briefly, formaldehyde (16% concentrate stock methanol-free, Cell Signaling Technology, 12606) was directly added to the media to a final concentration of 1% and incubated for 15 min. 125 mM Glycine (final concentration) was added to quench the reaction for 5 min. Cells were spun down and pellet was washed twice with PBS at 4°C. Protease (Millipore Sigma, 11836170001) inhibitors were added to all the buffers. The cell pellet was lysed on ice in cell lysis buffer (5 mM Pipes pH 8, 85 mM KCl, 0.5% NP-40). Pellet was centrifuged for 5 min at 1,700 g at 4°C. Pellet was resuspended with 0.5 mL of MNase digestion buffer (10 mm Tris pH 7.4, 15 mM NaCl, 60 mM KCl, 0.15mM Spermine, 0.5mM Spermidine) containing MNase (NEB, M0247S) to final concentration of final concentration of 20 U/µL and incubated at 37°C for 15 min. MNase was inactivated by the addition of EDTA (final concentration of 50mM) and brief incubation on ice. Pellet was centrifuged for 1 min at 16000g at 4°C. Pellet was resuspended in 0.5mL of nuclear lysis buffer (50mM Tris pH 8, 10mM EDTA, 0.5% SDS). Sonication was performed with a sonic Dismembranator Model 500 (Fisher Scientific) with the following parameters: amplitude 30%, 5 seconds on, 15 seconds off, 16 cycles. Sheared chromatin was centrifuged at 10,000 g for 15 min at 4°C. 15 µL of samples were de-crosslinked overnight at 65°C, size distribution was checked in 1% agarose gel and chromatin concentration measured using Nanodrop™. 5-10 µg of chromatin was used for each IP for RNAPII NTD and pSer2 RNAPII ChIP assay, and 100 µg of chromatin was used for each IP for CDK9 ChIP assay. ChIP spike in normalization for the RNAPII ChIP assays was performed using Drosophila-specific chromatin and *Drosophila*-specific histone variant (H2Av) antibody[119]. 50 ng of Drosophila S2 sheared crosslinked-chromatin (Active Motif, 53083) was added to 10 µg of chromatin as spike-in control. Chromatin was diluted with IP buffer (16.7mM Tris pH 8, 167 mM NaCl, 1.2 mM EDTA, 0.01% SDS, 1.1% TritonX-100) to obtain a 0.1% final concentration of SDS. 5% of diluted chromatin was kept as

input at 4°C. 10 µg of RNAPII NTD (D8L4Y) antibody (Cell Signaling Technology, 14958), 10 µg of pSer2 RNAPII (E1Z3G) antibody (Cell Signaling, 13499)), 10µg of CDK9 (C12F7) antibody (Cell Signaling Technology, 2316), or 1 µg of histoneH2Av (Active Motif, 39715) was mixed with diluted chromatin and subjected to IP performed on a rotating wheel at 4°C overnight. Pierce Protein A/G Magnetic beads (ThermoScientific, 78609) were then added to the chromatin-antibody mixture and 4°C rotation continued for 3 more hours. Beads were washed once with wash buffer 1 (0.1% SDS, 1% TritonX-100, 2mM EDTA, 20 mM Tris pH 8, 150 mM NaCl), four times with wash buffer 2 (0.1% SDS, 1% TritonX-100, 2mM EDTA, 20 mM Tris pH 8, 500mM NaCl), once with wash buffer 3 (0.25M LiCl, 1% NP-40, 1 mM EDTA, 10 mM Tris pH 8), once with wash buffer 4 (10 mM Tris pH 8, 1 mM EDTA). Immuno-bound chromatin was eluted at 55°C for 1 hour with elution buffer (50 mM Tris pH 8, 10 mM EDTA, 1%SDS, proteinase K 7 µg/mL) and de-crosslinked overnight at 65°C. DNA was extracted with one volume phenol:chloroform:isoamyl alcohol 25:24:1 (Sigma-Aldrich, P2069) and precipitated for 60 min at −20°C with sodium acetate and ethanol. Pellet was washed with 70% ethanol and resuspended in TE buffer. DNA quality and size distribution were checked on Fragment Analyzer. 10 ng of DNA was used for library preparation according to NEBNext® Ultra II DNA Library Prep Kit (NEB, E7645S). Purity and size distribution of the libraries were estimated using Fragment Analyzer. Size-selected libraries were sequenced on Illumina NovaSeq S4 (paired-end 2x150) to average depth of 50 million reads.

GRO-seq:

GRO-seq was performed as previously described[24] with minor modifications. 20 million cells from two biological replicates were used for each GRO-seq assay. Briefly, cells were washed twice with ice-cold PBS before adding swelling buffer (10 mM Tris-HCL pH 7.5, 2mM MgCl2, 3 mM CaCl2, 2U/ml Superbase-in (Invitrogen)). Cells were swelled for 5 min on ice, washed with swelling buffer + 10% glycerol and then lysed in lysis buffer (10 mM Tris- HCL pH 7.5, 2 mM MgCl2, 3 mM CaCl2, 10% glycerol, 1%l Igepal (NP-40), 2 U/ml Superasein) to isolate nuclei. Nuclei were washed twice with lysis buffer and resuspended in freezing buffer (40% glycerol, 5 mM MgCl2, 0.1 mM 0.5M EDTA, 50 mM Tris-HCL pH 8.3) to a concentration of $2x10^7$ nuclei per 100 mL. Nuclei were then frozen in dry ice and stored at -80°C until further use. Nuclei were thawed on ice just prior to the nuclear run-on. An equal volume of pre-warmed nuclear run-on reaction buffer (10 mM Tris-HCl pH 8, 5 mM MgCl2, 300 mM KCl, 1 mM DTT, 500 mM ATP, 500 mM GTP, 500 mM 4-thio-UTP, 2 mM CTP, 200 U/ml Superasein, 1% Sarkosyl (N-Laurylsarcosine sodium salt solution) was added and incubated for 7 min at 30°C for the nuclear run-on. Nuclear run-on RNA was extracted with TRIzol LS reagent (Invitrogen) following the manufacturer's instructions and ethanol precipitated. NRO-RNA was resuspended in water and concentration was determined with Nanodrop™. Up to 150 mg of RNA was transfer to a new tube and 5% (w/w) of Drosophila S2 spike-in RNA was added. RNA was then chemically fragmented

by mixing with 2X RNA fragmentation buffer (150 mM Tris pH 7.4, 225 mM KCl, 9 mM $MgCl_2$) and incubated at 94°C for 3 min 30 sec, before placing it back on ice and adding EDTA (to a final concentration of 50 mM) and incubating on ice for a further 2 minutes. Fragmentation efficiency was analyzed by running fragmented and unfragmented RNA on Agilent 2200 TapeStation using High Sensitivity RNA ScreenTapes following manufacturer's instructions. Fragmented RNA was incubated in Biotinylation Solution (20 mM Tris pH 7.5, 2 mM EDTA pH 8.0, 40% dimethylformamide, 200 mg/ml EZ-link HPDP Biotin (Thermo Scientific, 21341)) for 2h in the dark at 25°C, 800 rpm. After ethanol precipitation, the biotinylated RNA was resuspended in water and biotinylated-RNA was separated with M280 Streptavidin Dynabeads (Invitrogen). 100 ul/sample of beads were washed twice with 2 volumes of freshly prepared wash buffer (100 mM Tris pH 7.5, 10 mM EDTA pH 8.0, 1M NaCl, 0.1% (v/v) Tween-20) and resuspended in 1 volume of wash buffer and added to the biotinylated-RNA. After 15 min in rotation at 4°C, beads were washed three times with wash buffer pre-warmed at 65°C and three times with a room temperature wash buffer. 4-S-UTP containing RNA was eluted in 100 mM DTT buffer and purified with RNA Clean and Purification kit (Zymo Research, R1013) with in-column DNAse reaction to eliminate traces of genomic DNA. The eluted RNA was quantified with Qubit High Sensitivity Assay kit (Invitrogen) and used to produce barcoded cDNA sequencing libraries using the SMARTer® Universal Low Input RNA Kit (Takara, 634938). Libraries were sequenced on Illumina NovaSeq S4 (paired-end 2x150) to average depth of 50 million reads per sample.

TT-TL seq:

TT-TL-seq was performed as described[27], with a few modifications (primarily using HDPD biotinylation instead of MTSEA-biotin). About 15 million cells per sample was labeled with s^4U (by adding it to culture media at final concentration of 500 μM) for 5 minutes. Cells were harvested and washed in ice cold PBS and RNA extracted by Trizol and chloroform extraction, ethanol precipitation (1 mM DTT added to prevent oxidation of the s^4U RNA). Total RNA was resuspended and treated with TURBO DNase, then extracted with acidic phenol:chloroform:isoamyl alcohol and again ethanol precipitated. RNA was resuspended in water and concentration was determined with Nanodrop™. Up to 150 mg of RNA was transferred to a new tube and 5% (w/w) of Drosophila S2 spike-in RNA was added. RNA was then chemically fragmented by mixing with 2x RNA fragmentation buffer (150 mM Tris pH 7.4, 225 mM KCl, 9 mM $MgCl_2$) and heating to 94°C for 3 min 30 sec, before placing it back on ice and adding EDTA (to a final concentration of 50 mM) and incubating on ice for a further 2 minutes. Fragmentation efficiency was analyzed by running fragmented and unfragmented RNA on Agilent 2200 TapeStation using High Sensitivity RNA ScreenTapes following manufacturer's instructions. Fragmented RNA was incubated in Biotinylation Solution (20 mM Tris pH 7.5, 2 mM EDTA pH 8.0, 40% dimethylformamide, 200 mg/ml EZ-link HPDP Biotin (Thermo Scientific, 21341)) for 2h in the dark at 25°C, 800 rpm. After chloroform: isoamyl alcohol (24:1) extraction followed by purification (Qiagen

RNeasy Mini Kit, 74104), the biotinylated RNA was resuspended in water and biotinylated-RNA was isolated using Dynabeads MyOne Streptavidin C1 magnetic beads (Invitrogen), as previously described[27]. s^4U-biotinylated RNA was eluted from the beads by cleaving the HPDP Biotin in elution buffer (100mM DTT, 20 mM DEPC, pH 7.4, 1 mM EDTA, 100 mM NaCl, 0.05% Tween-20) and cleaned up using 1x RNAcleanXP beads (Beckman Coulter, A63987). For time lapse chemistry, 20 ul of isolated s^4U-enriched RNA was added to a mixture of 0.835 µl 3M sodium acetate pH 5.2, 0.2 µl 500 mM EDTA, 1.365 µl RNAse free water, and 1.3 µl trifluoroethylamine. 1.3 µl of $NaIO_4$ (200 mM) was then added dropwise, and the reaction mixture was incubated for 1 h at 45 °C. 1x volume of RNA clean beads were then added and purified as described above. 2 µl of reducing master mix (58 µl of water, 10 µl 1M DTT, 10 µl 1M Tris pH 7.4, 2 µl 0.5mM EDTA, 20 µl 5M NaCl) was added to each sample for a final volume of 20 µl, mixed, and incubated in thermocycler at 37C for 30 min. The mixture was then mixed with 1x RNAclean beads and RNA purified, as described above. The eluted RNA was quantified with Qubit High Sensitivity Assay kit (Invitrogen) and used to produce barcoded cDNA sequencing libraries using the SMARTer® Universal Low Input RNA Kit (Takara, 634938). Validation of TT-TL procedure was performed by qPCR (TableS7 for primer sequences). Libraries were sequenced on Illumina NovaSeq S4 (paired-end 2x150) to average depth of 50 million reads per sample.

CUT&RUN seq:

CUT&RUN was performed as previously described[120]. Briefly, 0.5 million K562 cells or 25,000 human CD34+ cells were harvested, washed, and bound to activated Concanavalin A-coated magnetic beads (activated in 20 mM HEPES, pH 7.9, 10 mM KCl,1 mM CaCl2,1mM MnCl2), then permeabilized with Wash buffer (20 mM HEPES, pH7.5, 150 mM NaCl, 0.5 mM spermidine and Roche protease inhibitor tablet) containing 0.05% Digitonin (Dig-Wash). The bead-cell slurry was incubated with 0.5 µl antibody (H3K4me3 (Cell Signaling Technology, 9751), H3K4me3 (Cell Signaling Technology, 5326), H3K27me3 (Cell Signaling Technology, 9733), H3K4me3 (Cell Signaling Technology, 9751), H3K27ac (Cell Signaling Technology, 8173), pSer2 RNAPII (Cell Signaling Technology, 13499), Rabbit IgG (Cell Signaling Technology, 2729)) in a 50 µL antibody buffer ((20 mM HEPES, pH 7.5 , 150 mM NaCl, 0.5 mM Spermidine, 0.01% digitonin, 2 mM EDTA, protease inhibitor) at 4°C overnight on a nutator. After 3 washes in 1 ml Dig-wash, beads were resuspended in 50 µl digitonin wash, mixed with 2.5 µl 20x pAG/MNase (EpiCypher, 151016) and incubated for 15 min at room temperature. After two washes in Dig-wash, beads were resuspended in 150 µl digitonin wash. Tubes were chilled to 0°C and 1 µl 100 mM CaCl2 was added while gently vortexing. Tubes were nutated at 4°C for 2 hours followed by addition of 33 µl EGTA-STOP buffer (340 mM NaCl, 20 mM EDTA, 4 mM EGTA, 50 µg/mL RNase A, 50 µg/mL Glycogen with E.coli spike in control (Epicypher, 18-1401)). Beads were incubated at 37°C for 30 min, replaced on a magnet stand and the liquid was removed to a fresh tube and DNA was extracted. DNA quality and size distribution were checked on Fragment Analyzer. 10 ng of

DNA was used for library preparation according to NEBNext® Ultra II DNA Library Prep Kit (NEB, E7645S). Purity and size distribution of the libraries were estimated using Fragment Analyzer. Size-selected libraries were sequenced on Illumina NovaSeq S4 (paired-end 2x150). Size-selected libraries were sequenced on Illumina NovaSeq S4 (paired-end 2x150) to an average depth of 10 million reads per sample.

ATAC-seq:

ATAC seq was performed using the Omni-ATAC protocol[121], optimized for low cell numbers. 50,000 K562 cells or 5000 CD34+ cells (flow sorted) were harvested, washed, and resuspended in cold PBS. Washed cells were resuspended in 1 ml of cold ATAC-seq resuspension buffer (RSB; 10 mM Tris-HCl pH 7.4, 10 mM NaCl, and 3 mM $MgCl_2$ in water) and centrifuged at 500g for 5 min in a 4 °C fixed-angle centrifuge. Cell pellets were then resuspended in 50 µl of RSB containing 0.1% NP40, 0.1% Tween-20, and 0.01% digitonin by pipetting up and down three times. This cell lysis reaction was incubated on ice for 3 min. After lysis, 1 ml of RSB containing 0.1% Tween-20 (without NP-40 or digitonin) was added, and the tubes were inverted to mix. Nuclei were then centrifuged for 10 min at 500g at 4 °C. Supernatant was carefully removed and nuclei were resuspended in 50 µl of transposition mix (25 µl 2× TD buffer (20 mM Tris pH 7.6, 10 mM $MgCl_2$, 20% Dimethyl Formamide) 2.5 µl transposase (100 nM final), 16.5 µl PBS, 0.5 µl 1% digitonin, 0.5 µl 10% Tween-20, and 5 µl water) by pipetting up and down six times. Transposition reactions were incubated at 37 °C for 30 min in a thermomixer with shaking at 1,000 rpm. Reactions were cleaned up with the MinElute PCR purification kit (Qiagen, 28004). DNA quality and size distribution were checked on Fragment Analyzer. 10 ng of DNA was used for library preparation according to NEBNext® Ultra II DNA Library Prep Kit (NEB, E7645S). Purity and size distribution of the libraries were estimated using Fragment Analyzer. Size-selected libraries were sequenced on Illumina NovaSeq S4 (paired-end 2x150).

Nascent RNA-seq (long and short read):

Nascent RNA-seq for short- and long-read sequencing was performed on K562 cells (~60 million per biological replicate), as described in previously published protocols[28] [1], with minor modifications. To minimize RNA degradation, all steps were performed on ice, and all buffers contained 25 uM α-amanitin, 40 U/ml RNase inhibitor (Invitrogen), and 1x Roche protease inhibitor mix. Briefly, 20 million cells were rinsed once with PBS/1 mM EDTA, then lysed in 250 µL cytoplasmic lysis buffer (10 mM Tris-HCl pH 7.5, 0.05% NP40, 150 mM NaCl) by gently resuspending then incubating on ice for 5 minutes. Lysate was then layered on top of a 500 µL cushion of 24% sucrose in cytoplasmic lysis buffer and spun at 2,000 rpm for 10 min at 4°C. The supernatant (cytoplasm fraction) was removed, and the pellet (nuclei) were rinsed once with 500 µL PBS/1 mM EDTA. Nuclei were resuspended in 100 µL nuclear resuspension buffer (20 mM Tris-HCl pH 8.0, 75 mM NaCl, 0.5 mM EDTA, 0.85 mM DTT, 50% glycerol) by gentle flicking, then lysed by the addition of 100 µL nuclear lysis buffer (20 mM HEPES pH 7.5, 1 mM DTT, 7.5 mM $MgCl_2$, 0.2 mM EDTA, 0.3 M NaCl,

1 M Urea, 1% NP-40), vortexed for twice for 2 seconds each, and then incubated on ice for 3 min. Chromatin was pelleted by spinning at 14,000 rpm for 2 min at 4°C. The supernatant (nucleoplasm fraction) was removed, and the chromatin was rinsed once with PBS/1 mM EDTA. Chromatin was immediately dissolved in 100 µL PBS and 300 µL TRIzol Reagent (ThermoFisher). RNA was purified from chromatin pellets in TRIzol Reagent (ThermoFisher) using the RNeasy Mini kit (QIAGEN) according to the manufacturer's protocol, including the on-column DNase I digestion. For genome-wide nascent RNA-seq, samples were depleted three times of polyA(+) RNA using the Magnetic mRNA isolation kit (New England Biolabs), each time keeping the supernatant, then depleted of ribosomal RNA using oligo probe-based rRNA depletion[122]. For short read sequencing, the isolated polyA, rRNA depleted RNA was then fragmented with a Bioruptor UCD-200 for 1-5 cycles of 30 s ON / 30 s OFF, high settings. Fragmentation efficiency was analyzed by running fragmented and unfragmented RNA on Agilent 2200 TapeStation using High Sensitivity RNA ScreenTapes following manufacturer's instructions. The fragmented RNA was used to produce barcoded cDNA sequencing libraries using the SMARTer® Universal Low Input RNA Kit (Takara, 634938). Short-read libraries were sequenced on Illumina NovaSeq S4 (paired-end 2x150). For long read sequencing, a DNA adapter (TableS6) was ligated to 3′ ends of nascent RNA using the T4 RNA ligase kit (NEB) by mixing 50 pmol adapter with 300-600 ng nascent RNA. cDNA was generated from the adapter ligated RNA using the SMARTer PCR cDNA Synthesis Kit (Clontech), replacing the CDS Primer IIA with a custom primer complementary to the 3′ end adapter for first strand synthesis (TableS6). cDNA was amplified by 15 cycles of PCR using the Advantage 2 PCR Kit (Clontech) with template switching custom oligo (TableS6), cleaned up using a 1X volume of AMPure beads (Agencourt), then PacBio library preparation was performed by the Yale Center for Genome Analysis using the SMRTbell Template Prep Kit 1.0 (Pacific Biosciences). The library was sequenced on four RSII flowcells and four Sequel 1 flowcells.

DRIP-seq:

DRIP-seq was performed as described[68]. K562 cells (5×10^6) were lysed in 0.5% SDS/TE, pH 8.0 containing Proteinase K overnight at 37 °C. Total DNA was isolated with phenol/chloroform/isoamylalcohol extraction followed by standard ethanol precipitation. One-third of total DNA was fragmented by a cocktail of restriction enzymes (EcoRI, HindIII, BsrgI, SspI, XbaI) overnight at 37 °C. A negative control treated overnight with RNaseH1 (New England Biolabs) was included. Digested DNA was purified by phenol/chloroform/isoamyl alcohol extraction, ethanol precipitation and quantified by Nanodrop™. 4 µg of digested DNA were diluted in binding buffer (10 mM $NaPO_4$, pH 7.0; 0.14 M NaCl; 0.05% Triton X-100) and incubated with 10 µg of S9.6 antibody overnight at 4 °C on a rotator. DNA/antibody complexes were added for 2 hours at 4 °C to Agarose Protein-A/G beads (ThermoScientific, 20421) prewashed with a binding buffer. Immunoprecipitated DNA was eluted by incubating with an elution buffer (50 mM Tris pH 8.0; 10 mM EDTA; 0.5% SDS) containing proteinase K at 55 °C for 45 min on a rotator. The eluent was precipitated by phenol/chloroform/isoamylalcohol extraction and ethanol precipitation. Validation of the DRIP procedure was

performed by qPCR (TableS7 for primer sequences). The pulled down material and input DNA were then sonicated, size-selected, and ligated to Illumina barcoded adaptors, using TruSeq ChIP Sample Preparation Kit (Illumina) or ThruPLEX® DNA-seq Kit (Rubicon Genomics) for next-generation sequencing (NGS) on Illumina HiSeq 2500 platform. NEBNext® Ultra II Prep Kit (NEB, E7645S). Purity and size distribution of the libraries were estimated using Fragment Analyzer. Size-selected libraries were sequenced on Illumina NovaSeq S4 (paired-end 2x150) to an average depth of XXX million reads per sample.

RNA-seq:

RNA was purified, fragmented and size checked on a 2100 Bioanalyzer (Agilent). Libraries were built using TruSeq Stranded Total RNA Library Prep (Illumina) after rRNA depletion according to the manufacturer's protocol. RNA-seq libraries were prepared using the Illumina TruSeq Stranded mRNA Library Prep Kit. Paired-end RNA-seq were performed with Illumina NovaSeq S4 (paired-end 2x150) to an average depth of 100 million reads per sample.

Bioinformatic analysis:

For all next generation sequencing datasets, the quality of the paired-end sequencing data was assessed with FastQC[123]. ATAC-, ChIP-, DRIP-, GRO- seq, TT-TL-seq data and Drosophila S2 spike in (for RNAPII ChIP) were trimmed with Cutadapt v4.1[124] to remove adaptor sequences. The trimmed reads were aligned to the canonical version of human reference genome hg38 and Drosophila genome (UCSC, dm6) using specific aligners as indicated for each assay.

ChIP-seq. : Bowtie2[125]
GRO-seq : Bowtie2 and STAR[126] (for splicing analysis)
TT-TL-seq : Bowtie2 and STAR (for splicing analysis)
nSRS-seq : Bowtie2
CUT&RUN: Bowtie2
RNA-seq : STAR

Mapping quality in the Bam files was assessed with SAMtools[127]. Then, PCR duplicates were removed, and the resulting reads were extended to 150 bases toward the interior of the sequenced fragment and were normalized to total reads aligned (reads per kilobase per million, RPKM). In cases when stranded libraries were built (GRO-seq, RNA-seq), reads assigned to Watson and Crick strands using SAMtools. For assessing correlation between biological replicates to determine reproducibility, the paired-end sequencing reads were mapped to human genome using Bowtie2. Average read coverages were calculated by feeding resulting bam files into 'multiBamSummary' program (deepTools Version 3.5.1) which was subsequently plotted using 'PlotCorrelation' program (deepTools Version 3.5.1). Average coverages, RPKM-normalized bigwig files, and

metaplot images were generated using deepTools 2.0[128]. Tracks were visualized using IGV[129]and generated using Gviz[130]. Further analyses were done in R (http://www.R-project.org), with Bioconductor packages and ggplot2[131] for graphic representation. With regard to RNA-seq, counts were assigned to every gene as mentioned and RPKM normalized for each replicate. After entire population analysis, those with an RPKM average from both replicates >0.001 were considered as expressed genes. rMATS[132] package was used to identify alternative splicing events from RNA-seq data. For comparative purposes in nascent SRS-seq, GRO-seq and RNA-seq datasets, counts per gene were calculated using FeatureCounts[133] for each condition and replicate. The raw count values were then subjected to default library-size normalization and differential expression analysis using DESeq2[134].
For DRIP-seq and ATAC-seq data, Epic2[135] package was used to call differentially enriched domains in WT and SF3B1^{K700E}, using default parameters. Differentially DRIP-seq and ATAC peaks were determined as those with a $|\log_2$ fold change(L_2FC)$| > 0.58$ and adjusted p value < 0.05, calculated with DESeq2. Genome and gene annotation of peaks was performed using the HOMER v4.11 package[136] and genes retrieved from Gencode version 29 release (2018). Volcano plots were generated using ggplot2 on L_2FC and adjusted p values. Convergent gene-pairs were identified and retrieved from Gencode version 29 release (2018) using an in-house python script.

The analyses of replication fork directionality and replication initiation zones used the published OK-seq data[137] in K562 cells from the ENCODE database. Bamscale[70] was used to create RFD scores from the OK-seq data and scaled coverage tracks for visualization in genome browsers.
RFD values were measured for each R-loop peak to establish R-loop orientation with respect to the replication fork (by intersecting R loop peaks with replication forks using BEDTools[138]). For purposes of RFD collision analysis, assays involving RFD were performed only on those sites (R-loops) where $|RFD| > 0.5$, ensuring a high chance of collision in a specific orientation.

Long read sequencing analysis:

Long Read Sequencing analysis was performed as previously reported[1], utilizing Porechop v0.2.4 (https://github.com/rrwick/Porechop) to remove external adapters, Cutadapt for removing 3' adapters and Prinseq-lite[139] to remove PCR duplicates. Minimap2[140] was used to align to hg38 build. Custom scripts in Python and R were used to generate Co-SE and cumulative distribution of 3' RNAPII position.

CUT&RUN seq analysis:

Paired-end reads were aligned to the human reference genome hg38 using Bowtie2. Mapping quality in the Bam files was assessed with SAMtools. PCR duplicates were removed, and the resulting reads were normalized to total reads aligned (RPKM). Average coverages, RPKM-normalized bigwig files, and metaplot images were generated using deepTools 2.0. Tracks were generated using Gviz. Further analyses were done in R (http://www.R-project.org), with Bioconductor packages and ggplot2 for graphic representation. Peak calling was performed using SEACR package[141] using control IgG CUT&RUN bedgraph files to generate empiric thresholds.

TT-TL-seq analysis:

TT-TL-seq reads from SF3B1^{WT}, SF3B1^{K700E} and U2AF1^{WT}, U2AF1^{S34F} cells containing spiked-in RNA from *Drosophila* S2 cells were trimmed with Cutadapt to remove adaptor sequences. Trimmed reads were mapped to a combined hg38 and dm6 genome using Bowtie2. Bam files for uniquely mapped reads (SAM flag 83/163 or 99/147) were created, sorted and indexed using Samtools. To calculate normalization factors, reads aligning to the dm6 genome were filtered and counted with Samtools. RPKM-normalized, Drosophila-scaled bigwig files and metaplot images were generated using deepTools 2.0. To assess the TL-component in the TT-TL seq, unfiltered, trimmed reads were mapped to a combined hg38 and dm6 genome using HISAT-3N[142] with default parameters and U-to-C mutation calls. Normalized coverage tracks were generated using Gviz.

For accurate differential expression analysis using DESeq2, normalization factors for the sample libraries were calculated with FeatureCounts, using readcounts mapped to the dm6 genome in addition to library size normalization. Differential expression analysis of normalized read counts was performed in gene bodies (TSS+300bp to TES).

Sample	**DESeq2 normalization factor**
WT-1	0.66352
WT-2	0.8424
SF3B1^{K700E} 1	1.57728
SF3B1^{K700E} 2	1.16532

Clusterprofiler was used for gene ontology enrichment analysis of biological enrichments. Read-in and read-through values were generated for TT-TL-seq (reads with 1 mutation or more) as well as GRO-seq using ARTDeco[143].

Identification of promoter and enhancer analysis:

Genome annotation and reference genome sequences were downloaded from the UCSC Genome Browser. Only those RefSeq-validated transcripts with a GRO-seq signal at the region between TSS – 10 bp and TSS + 300 bp were selected. Transcripts that are ≤ 1 kb long and ≤ 1 kb from the nearest genes were removed. TSS of antisense transcription was defined by the highest GRO-seq signal in the antisense strand within 500 bp upstream and downstream of the sense TSS. The highest GRO-seq signal within the region from +20 to +300 bp downstream of TSS was defined as the pausing site. The promoter was defined as the region between TSS – 100 bp and TSS + 300 bp.

For putative enhancer identification, distal H3K27ac peaks (determined from K562 CUT&RUN data), without overlap of the region from 2kb upstream and downstream of the protein coding genes, were defined as enhancers and the strand having higher GRO signal was defined as the sense direction. The pausing site of the enhancer was defined as described above, for promoters.

Chromatin state analysis:

ChromHMM[86] was used with default parameters to derive genome-wide chromatin state maps for all cell types. Input ChIP-seq, ATAC-seq, DRIP-seq, and CUT&RUN data were binarized with ChromHMM's BinarizeBed method using a p-value cutoff of 10^{-4}. Chromatin state models were learned jointly on all the marks from $SF3B1^{WT}$ and $SF3B1^{K700E}$ from 10 to 15 states. A model with 14 states was chosen for detailed analysis. Chromatin state annotations of $SF3B1^{WT}$ and $SF3B1^{K700E}$ were produced subsequently by applying this model to the original binarized chromatin data from these cell types. To determine important chromatin state changes between WT and $SF3B1^{K700E}$, the chromatin state annotations were intersected. In each case, the number of 200bp bins that is occupied by each of the 14 by 14 possible chromatin state transitions. To calculate enrichment scores, this number was divided by the expected number of such bins assuming a null model that treats the two chromatin states involved in each transition as being independently distributed.

For promoter state analysis, promoters were defined as regions -1Kb to +1Kb proximal to TSS. Every RefSeq gene was assigned to one chromatin state based on the state call on the gene's TSS in WT and $SF3B1^{K700E}$ cell types by using the 14 state ChromHMM genome annotation output.

Unsupervised clustering of Alternative Splicing (AS) events:

RNA-seq files were aligned to GRCh8 by STAR and bam files were used to determine percentage spliced in (PSI) for each sample using rMATS. Splicing events were aggregated and those with at least PSI of 0.1 in 10%

samples (22 of 217 total) were used for unsupervised clustering using ComplexHeatmap package[144] in R. To optimize performance, 10,000 random events (from total 137,435 that met PSI criteria) were chosen for clustering. Default parameters (Pearson distance method) were used to cluster by rows (AS events). Annotations for disease status (MDS/AML or normal) and splicing status of SF3B1, U2AF1, SRSF2 and ZRSR2 were provided.

shRNA screen:

~8 million K562 cells, per replicate, were transduced with a HIV-based lentiviral RNAi library[145] containing a total of 2343 shRNAs (targeting 357 genes) at a MOI of 0.2 and a target representation of 3000 cellular integrations for each shRNA. After transduction, cells were selected for 4 days in puromycin, then grown in doxycycline, and cultured for ~12 population doublings (24 days). During this time, cells were passaged every 4 days. Cells were collected immediately after puromycin selection (day 0) and subsequently on days 8, 16 and 24 after doxycycline induction (2 million cells each). The QIAamp DNA Blood Midi kit (Qiagen) was used to extract genomic DNA from cell pellets followed by ethanol precipitation to clean and concentrate. Half-hairpin sequences were amplified from the genomic DNA (common forward primers 5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTCTTGTGGAAAGGACGA-3′ and reverse primer 5′-CAAGCAGAAGACGGCATACGAGATNNNNNNNNGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTCTACTATTCTTTCCCCTGCACTGT -3′) using Q5 High-Fidelity 2x Master Mix (NEB). PCR reactions were cleaned up using Agencourt AMPure XP beads (Beckman Coulter) followed by measurement of the concentration of the target amplicon using the Agilent 4200 TapeStation. DNA samples were indexed using the Nextera XT kit (Illumina). Indexed and size-selected samples were quantified using the KAPA Illumina Library Quantification kit (KAPA Biosystems) and pooled at equal concentration prior to sequencing using the Illumina NovaSeq S4 (paired-end 2x150).

shRNA data processing and analysis:

A custom reference library of the shRNA library was created for Bowtie2. Adaptor sequences were removed from the reads using Cutadapt, followed by alignment to the custom reference library using Bowtie 2. The number of reads mapping to each shRNA in each sample was then extracted from the SAM files and Mageck2[146] was used to determine the fold-change in abundance of each shRNA between the D0 time point and later time points (D8, D17, D24) in both WT and SF3B1^{K700E}. shRNA-level fold-change estimates were combined to gene-level estimates by averaging the sum of individual fold-changes of all significant shRNAs (adjusted p value < 0.05 and $|FC| > 0.5$) mapped to a given gene. Gene-level p values were calculated by

combining the individual p values of all significant shRNAs mapping to a given gene using Fisher's method. Gene-level weighted growth effects were calculated by multiplying the average gene-level fold change by the -$\log_{10}$ (gene-level p value). Clusterprofiler R package[147] was used for gene ontology enrichment analysis of target gene enrichments. Analysis of over-represented gene pathways was performed using the clusterprofiler R package with the following parameters: OrgDb = 'org.Hs.eg.db', ont = "BP",pvalueCutoff = 0.05, qvalueCutoff = 0.10. Pathway over-representation for screen candidates was performed using STRINGdb v11.5[148].

Methylcellulose colony forming unit assay:

Mononuclear cells were isolated by Ficoll-Hypaque density gradient separation (Sigma Aldrich, Histopaque-1077) and cultured in methylcellulose media (Methocult GF#H4435, Stem Cell Technologies) per manufacturer's instructions. Similarly, FACS-sorted cKit+Sca-1+Lin- cells isolated from murine bone marrow aspirates were cultured in methylcellulose median (Methocult GF#M3434, Stem Cell Technologies). The following independent experiments were performed in technical duplicates for each setting: 0.05% DMSO and 5 uM OICR-9429 (SelleckChem, S7833) in 0.05%DMSO). After 14 days of incubation, erythroid or myeloid colonies were counted, representative pictures were taken, and cultures aspirated, washed in PBS and stained with antibodies for flow cytometry. To knockdown human HDAC2 and ING2 (and Hdac2, Ing2, Wdr5 in mouse), the FACS-isolated cells (were transduced with pLKO.1-based GFP-expressing scrambled, HDAC2, ING2 (scrambled, Hdac2, Ing2, Wdr5 in mouse) -targeting lentiviral vectors by spinfection at 1800 rpm for 90 minutes at 37°C in 8 µg/mL polybrene. After 6 hours of incubation in IMDM + 10% FBS at 37°C, stably transduced cells were cultured in methylcellulose media as detailed above. Colony counting was performed manually using STEMgrid-6 (StemCell technologies, #27000). Images of the well plates were taken using Keyence BZX-800 inverted fluorescence microscope. For the human samples, the same sequences previously validated in our shRNA rescue screen validation experiments were used to knockdown HDAC2 and ING2. Knockdown efficiency of the respective proteins was confirmed using Western Blots. The shRNA sequences used to knockdown HDAC2, ING2 in the human samples and Hdac2, Ing2, and Wdr5 in the murine LT-HSC cells are included in TableS4.

Statistics and reproducibility:

Statistical parameters including the number of biological replicates (*n*), s.e.m. and statistical significance are reported in the figure legends. All results presented in this article were obtained from a minimum of three independent biological replicates, except for genome-wide analyses and DNA combing. CUT&RUN-seq for Ser2P RNAPII and H3K4me3 in human CD34+ cells was performed in a minimum of 5 biological replicates, per condition. DNA combing, K562 DRIP-seq, K562 ATAC-seq, K562 nascent SRS-seq, K562 LRS, K562

ChIP-seq, K562 GRO-seq, K562 TT-TL seq, K562 RNA-seq, K562 H3K4me3-CUT&RUN-seq, K562 H3K27me3-CUT&RUN-seq, and murine LT-HSC-CUT&RUN-seq data resulted from two independent biological replicates for each condition. U2AF1 (WT and S34F) Ser2P RNAPII ChIP-seq, H3K4me1-CUT&RUN-seq, and H3K27ac-CUT&RUN-seq, shRNA and siRNA knockdown Ser2P RNAPII ChIP-seq data came from one biological replicate per condition. K562 GRO-seq (1 replicate), H3K36me3 ChIP-seq (1 replicate) and OK-seq (three replicates) data were retrieved from external database repositories.

For pair-wise comparison, Student's *t*-test was used unless otherwise specified. When data was presented as box-whisker plots, the Mann-Whitney U test was performed. Test details are indicated in the figure legends. For box plots, boxes and whiskers indicate the 25th to the 75th and the 5th to the 95th percentiles, respectively, and medians are indicated. Pearson's correlation was calculated for correlation analysis. In Venn diagrams, numbers represent genes or regions co-occurring between conditions. In screenshots from genome-wide experiments, scales are adjusted so that the background signal is low for better visualization of the results.

For IF (S9.6, γH2AX) and co-IF (γH2AX + BrdU, γH2AX + p-RPA, γH2AX + S9.6) and EdU experiments, >100 cells per replicate were measured. A minimum of 50 cells was also measured in each comet assay replicate. For DNA-combing analysis, at least 150 tracks per replicate were measured to determine fork velocity. In RNAPII NTD + PCNA PLA and PCNA + FANCD2 PLA plots, >100 total cells were represented from at least 30 cells per replicate.

Graphs were generated with the ggplot2 package of R. *P* values are indicated in the figures, and statistical tests applied are described in the figure legends. Analyzed samples were randomly chosen and data acquisition automatically performed by analysis software to ensure unbiased results.

Data Accession: All sequencing files used in this study are available from GEO with accession number GSE226003

Author Contributions: PCB and MMP conceived and planned the study. PCB, AKG performed the microscopy and NGS experiments. BD and ED performed the DNA fiber combing experiments. RR, NN, HL contributed to the in vitro splicing assay experiment. BD, ED performed the DNA combing studies. AO, GK contributed to the DNA replication studies experiments. JZ, MS contributed to and supervised the bioinformatic analysis for the TT-TL-seq experiment. NC, DK, AKV provided patient samples for the study. YN and SO provided RNA-seq files of some patient samples. PCB and MMP performed bioinformatic analysis and wrote the manuscript. KMN provided input on experimental design and analysis of long read

sequencing. KMN and AKV edited the manuscript. MMP supervised overall supervision for the project. All the authors discussed and approved the final version of the manuscript.

Acknowledgements: This work was funded in part by National Institutes of Health (R01 HL 133406 to MMP and KMN, R21 HL150642 to MMP and KMN, T32 CA233414 to PCB, NIDDK CCEH pilot to PCB), pilot awards from the Frederick A. DeLuca Foundation (to MMP and KMN), EvansMDS Young Investigator Award from the Edward P. Evans Foundation (to PCB). We thank the Yale Center for Genome Analysis (YCGA), Yale Center for Research Computing (YCRC), Yale Animal Resource Center (YARC), Yale Core Center of Excellence in Hematology (YCCEH, supported by NIDDK U54 DK106857) for the core resources used. Some of the patient samples used in the study were provided by The Yale Hematology Tissue bank. We thank Dr. Carl Walkley (St. Vincent's, Melbourne, Australia) and Dr. Alex C. Minella for critical insights and discussions.

FIGURE LEGENDS:

Figure 1: **Genome-wide analysis of RNAPII redistribution due to SF3B1^{K700E}** (A) Inducible *SF3B1* mutant isogenic cell line model system. Time course experiment collects cells at 96 hours after Doxycycline induction of Cre-recombinase expression. Cre-recombinase expression results in loxP excision of the 'floxed' intron trap cassette and expression of the now-in-frame modified allele. Mutation in the exon 15 of *SF3B1* gene was generated in K562 cells using AAV-intron trap and CRISPR-Cas9. The mutation is 'A' to 'G' which results in addition of Glutamic acid in place of Lysine at 700 amino acid position in *SF3B1*. Wild-type K562 cells are subject to the exact same AAV-intron trap CRISPR/Cas9-mediated genome manipulation except that the wild-type cells have synonymous change (A to A). (B) A metagene plot showing the Drosophila spike-in-normalized ChIP-seq levels of RNAPII NTD over 6694 expressed genes in WT and SF3B1^{K700E} cells. Solid lines represent the averaged signal, and the shaded area represents the upper and lower signal limits of the individual replicates (n =2, WT; n =2, SF3B1^{K700E}). Read densities are calculated as RPKM. (C) Representative genome browser track examples of NTD RNAPII ChIP-seq in K562 cells WT (blue) versus SF3B1^{K700E} (orange), at 4 days post doxycycline induction, for the indicated genes (n=2 biological samples, each). The x-axis indicates the chromosome position, and the y-axis represents normalized read density in RPKM. (D) The cumulative distribution function (CDF) plot of the TR distribution (calculated as an average of two replicates) in WT and SF3B1^{K700E} cells (n = 6694 genes). Significance determined using two-sample K.S. test. Inset: A schematic diagram depicting the calculation of the TR as the ratio of RNAPII read density in the TSS to TSS+300 bp (Promoter) versus the signal in the TSS+300 bp to transcription end site (Gene body)

of each transcript. (E) Meta plots showing the log_2 fold change signal of RNAPII Ser2P ChIP-seq IP data (normalized to input) across the annotated gene transcripts genes (n = 6694 genes) WT and SF3B1^{K700E} K562 cells. Solid lines represent the averaged signal, and the shaded area represents the upper and lower signal limits among the individual replicates (n =2, WT; n =2, SF3B1^{K700E}). Read densities are calculated as RPKM. (F) Heat maps depict change in RNAPII NTD and Ser2P CTD signals upon SF3B1^{K700E} expression in K562 cells (i.e. (average RNAPII enrichment in WT) – (average RNAPII enrichment in SF3B1^{K700E})). Unsupervised K clustering based on RNAPII NTD and RNAPII PS2 changes separates active genes into three distinct clusters (number of clusters determined using R-NbClust package). Clustering calculated based on change in ChIP-seq signals (WT vs SF3B1^{K700E}) from −1 kb upstream of TSS (transcription start site) to 2 kb downstream of the TES (transcription end site). Genes in clusters 1 are characterized by highest redistribution of RNAPII densities into gene bodies versus genes in cluster 3 which have the least redistribution in RNAPII densities into the gene body. The heatmap patterns are also reflected in the corresponding 'TR' cumulative distribution graphs for each of the 3 clusters (right panel). Cluster I = 4289 genes, Cluster 2 = 1317 genes, and cluster 3 = 820 genes. (G) Box-whisker plots showing the distribution of gene length (top) and intron number (bottom) for the three gene clusters. Dots represent data for individual genes (two-tailed Mann–Whitney *U*-test. ***$P = 0.0001$, *$P = 0.02$, $^{ns}P > 0.05$). Gene length (in bp) and Intron number have been log-transformed for better visualization. (H) Venn diagrams showing distribution of intronless genes across the 3 gene clusters (total, n=6994 gene dataset; total intron less genes = 200). (I) SF3B1^{K700E}/WT ratio RNAPII NTD ChIP heatmap with static binning spanning the entire length of all expressed genes (n = 6694) up to 50 kb ordered by increasing gene length. Read densities are in RPKM. Corresponding metaplot of density ratio shown on the top of the heatmap. (J) Averaged RNAPII enrichment profiles for all intron-containing genes in SF3B1^{K700E} cells compared to WT (SF3B1^{K700E}/WT density ratio, n = 5118 genes). Genes were divided into two bins representing first exon and first intron, as indicated on the X-axis. Y-axis shows the density ratio in RPKM.

Supplemental Figure 1: **Genome-wide analysis of RNAPII redistribution due to SF3B1**^{K700E} (A) Correlation between RNAP II NTD ChIP-seq samples. Values in boxes represent Pearson's correlation coefficients between corresponding samples (WT and SF3B1^{K700E} K562, n = 2 biological replicates each). (B) Values in boxes represent Pearson's correlation coefficients between corresponding samples (WT and SF3B1^{K700E} K562, n = 2 biological replicates each). (C) Individual Ser2P RNAPII ChIP-seq gene tracks at *PEX2 and SERBP1*— in K562 cells WT versus SF3B1^{K700E}, at 4 days post doxycycline induction. The x-axis indicates the chromosome position, and the y-axis represents normalized read density in RPKM. (D) Meta plots comparing the RNAPII NTD ChIP signal (Top panel) log_2 fold change signal of RNAPII Ser2P ChIP-seq IP/input (Bottom panel) between WT and SF3B1^{K700E} K562 cells, for the 3 gene clusters (n=2 biological replicates, each). Solid lines represent the averaged signal, and the shaded area represents the upper and lower signal limits of the individual replicates (n =2, WT; n =2, SF3B1^{K700E}). Read densities are calculated as RPKM. (E) Box-whisker plots showing the distribution of neighboring gene distance (left) and gene expression in FPKM

(right) for the three gene clusters. Dots represent data for individual genes (two-tailed Mann–Whitney *U*-test. ***$P = 0.0001$, *$P = 0.02$, $^{ns}P > 0.05$). Neighboring gene distance and gene expression have been log-transformed for better visualization. (F) Top panel: The cumulative distribution function (CDF) plot of the TR distribution (calculated as an average of two replicates) in WT and SF3B1^{K700E} cells depending on gene length. Gene lengths- <=1kb: 215 genes, 1-5kb: 599 genes, 5-10kb: 823 genes, 10-20kb: 1078 genes, 20-50kb:1638 genes, >50kb: 1480 genes. Bottom panel: The CDF plot of the TR distribution (calculated as an average of two replicates) in WT and SF3B1^{K700E} cells depending on gene intron content/number. Intronless: 191 genes, single intron: 466 genes, 2-4 introns: 1503 genes, 4-6 introns:1537 genes, 6-10 introns: 1485 genes, >10 introns: 597 genes. (G) Top: Pearson correlation scatter plot comparing gene length (log(10) transformed gene length) and percentage change in traveling ratio (WT_{TR}-SF3B1$^{K700E}{}_{TR}$/WT_{TR}). Bottom: Pearson correlation scatter plot comparing gene intron number ((log(2)transformed intron number)+ 1) and percentage change in traveling ratio (WT_{TR}-SF3B1$^{K700E}{}_{TR}$/WT_{TR}). (H) Read coverage in the intronic region of interest (From left to right: first, middle, and final introns; n =5118, each) in multi-intronic genes (greater than 2 introns) from the 6694 gene dataset. Coverage is normalized to the position 100 nt upstream of each 5'SS. Orange line indicates SF3B1^{K700E} and the blue line indicates WT.

Figure2: **Determination SF3B1^{K700E} induced changes in RNAPII transcription kinetics using nascent transcriptome assays** (A) Metaplot analysis showing GRO-seq densities in WT- and SF3B1^{K700E}-mutant K562 cells (TSS -10kb to TES +10kb regions) (n = 6694 genes) (n =2 biological replicates, each). Sense reads shown in positive orientation and antisense reads shown in negative orientation. Abbreviations used: TSS-transcription start site, TES-transcription end site. Read densities are calculated as RPKM. (B) Representative genome browser track examples of GRO-seq in K562 cells WT (blue) versus SF3B1^{K700E} (orange), at 4 days post doxycycline induction, for the indicated genes (n=2 biological samples, each). The x-axis indicates the chromosome position, and the y-axis represents RPKM. (C) The cumulative distribution function plot of the TR distribution, for GRO-seq, calculated as an average of two replicates) in WT and SF3B1^{K700E} cells (n = 6694 genes). Significance determined using two-sample K.S. test (p=0.000313). (D) GRO-seq read coverage in the introns (n =96881) and exons (n =96881), in WT and SF3B1^{K700E} cells. Coverage is normalized to the position 100 nucleotides (nt) upstream of the 5'SS (for intron) or 5' start (for exon). Orange line indicates SF3B1^{K700E,} and the blue line indicates WT. (E) Box-whisker plots comparing the fold changes (SF3B1^{K700E} versus WT) in GRO-seq gene body (TSS +300 bp to TTS) density and mRNA-seq abundance upon SF3B1^{K700E} expression. Fold changes in GRO-seq densities are shown for genes significantly up-regulated in SF3B1^{K700E} cells by > |0.5| log2fold in mRNA-seq ($n = 327$, $P_{adj} < 0.05$) and for genes significantly down-regulated in SF3B1^{K700E} cells by > |0.5| log2fold in mRNA-seq ($n = 405$ genes, $P_{adj} < 0.05$) (right), at 4 days post-doxycycline induction. (F) Genome browser tracks of TT-TL seq data for *SERBP1* and *SERTAD1* genes provide examples of results from the WT and SF3B1 ^{K700E} cells after normalization to Drosophila spike-in

control. (G) Metaplot showing Drosophila spike-in normalized TT-seq densities in WT- and SF3B1^{K700E}-mutant K562 cells (TSS -3kb to TES +3kb regions) (n = 6694 genes, 2 biological replicates each). Solid lines represent the averaged signal, and the shaded area represents upper and lower signal limits among the individual replicates. (H) Hierarchical clustering demonstrating similarities in the log2 fold changes of gene body density (TSS+300 to TTS) levels by GRO-seq, nascent RNA-seq, and TT-TL-seq without corresponding changes in gene expression as determined by total RNA-seq. (I) The CDF plot of the median 3'end RNAPII position (plotted as a ratio with respect to the gene length) as determined by the LRS 3'read ends, in WT and SF3B1^{K700E} (n = 6694 genes). Significance determined using two-sample K.S. test. (J) Global analysis comparing CoSE for introns in SF3B1^{K700E} versus those in WT. Analysis only included introns covered by at least 10 reads and for genes greater than 0.8kb (n = 367105 introns in SF3B1^{K700E}, n = 433596 introns in WT). Significance tested by two tailed Mann Whitney U test. (K) Ratio of spliced reads over total unspliced and spliced reads in WT and SFB3B1^{K700E} in GRO-seq. All the introns in the selected 5118 isoforms (from the 6694 gene dataset) were considered. SS-splice sites. Boxes represent the range between the first and third quartiles for each condition. Black horizontal lines in boxes show the median. n=2 biologically independent experiments. (L) CoSE in SF3B1^{K700E} compared to CoSE in WT for introns with statistically significant loss of intron retention in SF3B1^{K700E}, based on mRNA-seq (SF3B1^{K700E} vs WT) performed at 8 days. Analysis only included introns covered by at least 10 reads (n = 83 introns in SF3B1^{K700E} and WT). Significance tested by Mann Whitney U test

Supplemental Figure 2: **Determination SF3B1^{K700E} induced changes in RNAPII transcription kinetics using nascent transcriptome assays** (A) Scatter plots depicting correlation between GRO-seq samples in WT and SF3B1^{K700E} cells (n=2 replicates, each). Values in boxes represent Pearson's correlation coefficients between corresponding samples (*n* = 2 biological replicates). (WT, r=0.986; SF3B1^{K700E}, r=0.991). Values expressed Log2(gene-expression + 1). (B) CDF plots of TR from GRO-seq for 'ChIP-seq-derived' maximally affected (Cluster I), moderately affected (Cluster 2) and less affected (Cluster 3) genes in WT (blue) versus SF3B1^{K700E} K562 (orange) cells. Cluster I = 4289 genes, Cluster 2 = 1317 genes, and cluster 3 = 820 genes. (C) Scatter plots depicting correlation between TT-TL seq samples in WT (n=2) and SF3B1^{K700E} (n=2). Values in boxes represent Spearman's correlation coefficients between corresponding samples (*n* = 2 biological replicates). (WT, spearman=0.93; SF3B1^{K700E},spearman=0.99). Values expressed Log2(gene-expression + 1). (D) Scatter plots depicting correlation between nascent short read RNA-seq samples in WT and SF3B1^{K700E} cells. Values in boxes represent Spearman's correlation coefficients between corresponding samples (*n* = 2 biological replicates). Values expressed Log2(gene-expression + 1). (E) TT-seq RT–qPCR analyses showing relative fold enrichment of pre-mRNA (intron-exon spanning primers) and spliced/mature mRNA (exon-exon primers) for *CDKN2B, MYC, MIR17HG,* and *JUN* (intronless) genes in WT and SF3B1^{K700E} cells, 4 days post doxycycline induction. MYC spliced/mature mRNA (IP/input) was set as 1. WT (*n* = 2; WT-R1 and WT-R2) and SF3B1^{K700E} (n=2; SF3B1^{K700E}-R1 and SF3B1^{K700E}-R2). Data plotted are for the mean of two technical

replicates per condition. (F) Minus-average plot plotting the log2 fold change (n = 6694 genes) on the y axis and log(e)-transformed normalized mean counts on the x axis. Significantly upregulated and downregulated genes (L2FC > |0.5|, P_{adj}<0.05) in SF3B1^{K700E} (vs WT) are shown in red, above and below the zero line of the y axis, respectively. (G) Scatter plot comparing $WT_{TT\text{-}seq}$ RPK (Log-transformed; plotted on Y axis) and ratio of SF3B1$^{K700E}{}_{TT\text{-}seq}$ to $WT_{TT\text{-}seq}$ RPK (plotted on X axis). Spearman correlation of -0.06. (H) Meta plot showing the $\log_2$ fold change signal of RNAPII Ser2P ChIP-seq IP data (normalized to input) across the annotated gene transcripts genes (n = 6694 genes) in WT (blue) and U2AF1^{S34F} (orange) K562 cells (n =1, WT; n =1, U2AF1^{S34F}). Read densities are calculated as RPKM. Abbreviations used: TSS-transcription start site, TES-transcription end site (I) Metaplot showing Drosophila spike-in normalized TT-seq densities in WT (blue) and U2AF1-mutant (orange) K562 cells (TSS -5kb to TES +5kb regions) (n = 6694 genes, 2 biological replicates each). Solid lines represent the averaged signal, and the shaded area represents upper and lower signal limits among the individual replicates. (J) Read length distribution of PacBio long-reads. (K) Read depth distribution of PacBio long-reads. (L) LRS data are shown for WT (top) and SF3B1^{K700E} (bottom) K562 cells for the *FTL* gene. Reads are aligned to the genome and ordered by 3′ end positions. Each horizontal row represents one read. Each transcript was categorized and colored according to its splicing status, which can be either completely spliced (grey), partially spliced (purple), all unspliced (orange), or "NA" (green) (i.e., transcripts that did not span an entire intron or a 3′SS). 'n' is the total number of reads aligning to the selected area of the genome for a particular condition. Bar graphs at the far left (for both the upper and lower plots) indicate the fraction of the aligned intron-spanning reads that are partially spliced (purple), unspliced (orange), fully spliced (grey) for that gene. (M) The CDF plot of the average 3'end RNAPII position (as determined by the LRS 3'read ends) in WT and SF3B1^{K700E} in the 3 RNAPII ChIP-based gene clusters, shown in Figure 3E. Significance determined using two-sample K.S. test. (N) CoSE for introns in SF3B1^{K700E} compared to the CoSE for introns in WT separated by the position of intron (first, middle, and last introns from left to right) in the selected 5118 isoforms. Analysis only included introns covered by at least 10 reads. Significance tested by Mann Whitney U test. (O) Ratio of spliced reads over total unspliced and spliced reads in WT and SF3B1^{K700E} in GRO-seq at the 5'SS and 3'SS. All the introns for the first, middle, and last (from left to right) in the selected 5118 isoforms were considered. SS-splice sites. n=2 biologically independent experiments. (P) Scatter plot comparing CoSE under WT and SF3B1^{K700E} conditions (n=83 introns), related to Fig2L.

Figure 3: **Mechanisms underlying SF3B1^{K700E} induced transcription elongation defect** (A) The FLAG-tag immunoprecipitated proteins from the untransfected, FLAG-tagged *SF3B1^{WT}* or *SF3B1^{K700E}* transfected HEK293T total nuclear (left) and chromatin (right) extracts were analyzed by Western blots. Abbreviations used: UT: untransfected, WT: FLAG-SF3B1^{WT} transfected (B) Immunoblot analysis of the association of the indicated proteins, including splicing factors, in nucleoplasmic and chromatin fractions of WT and SF3B1^{K700E} K562 cell lysates, at 6 days post-doxycycline induction. (C) Native gel analyses of the kinetics of spliceosome

assembly on radiolabeled *AdML* pre-mRNA in HEK293T (*SF3B1*^{WT} and *SF3B1*^{K700E}) nuclear extracts. The identities of spliceosome complexes E and A are shown. Relative amounts of A and E complexes were quantified and plotted over time as ratio of A-to-E. (D) Meta plots showing the $\log_2$ fold change signal of RNAPII Ser2P ChIP-seq IP data (normalized to input) across the annotated gene transcripts genes (n = 6694 genes) in SF3B1^{K700E} K562 and WT stably integrated with Tet-PLKO-Puro^{R} lentiviral plasmid constructs expressing inducible shRNA targeting either control (shScrambled), DDX46 (shDDX46), and HTATSF1 (shHTATSF1). shRNA expression induced with doxycycline, and ChIP performed 4 days after induction. (E) SF3B1^{K700E} + empty vector (EV), SF3B1^{K700E} + HA-tagged DDX46 overexpression (DDX46-OE), and SF3B1^{K700E} + HA-tagged HTATSF1 (HTATSF1-OE) overexpression cells were induced with doxycycline (Day 0 at induction time point), and their cell growth plotted over a course of 12 days ($n = 2$, error bars indicate SEM). Significance assessed using two-tailed unpaired *t* test (***$P_{5\ days(SF3B1^{K700E}\ +\ EV\ vs\ SF3B1^{K700E}\ +\ HTATSF1\text{-}OE)} < 0.0001$, **$P_{12\ days(SF3B1^{K700E}\ +\ EV\ vs\ SF3B1^{K700E}\ +\ HTATSF1\text{-}OE)} < 0.001$). (F) A schematic of the proposed model for SF3B1^{K700E}-induced RNAPII transcriptional defect. According to the intron-loop model, following the A complex formation, the U4/U6/U5 tri-snRNP join, and the subsequent transition to the B complex displaces U1 snRNP, releasing the RNAPII from the pre-catalytic spliceosome. Decreased engagement of HTATSF1 with chromatin-bound SF3B1^{K700E} prevents the ATP-dependent splicing steps of early spliceosome formation from occurring, with subsequent lack of formation of B-complex formation, thereby not allowing the release of RNAPII from U1snRNP. (G) Metagene analysis showing the occupancies of CDK9 (IP/Input) as measured by ChIP-seq in WT and SF3B1^{K700E} cells (n=2 biological replicates, each). Solid lines represent the averaged signal, and the upper and lower limits of the shaded area represent upper and lower signal limits among the individual replicates. (H) Quantification of readthrough phenotypes in WT and SF3B1^{K700E} cells in GROseq. Distribution of read-through levels (log2 ratio of reads in read-through region vs. gene body) and read-in levels (log2 ratio of reads in read-in region vs. gene body) for top 1000 expressed genes.

Supplemental figure 3: **Mechanisms underlying SF3B1^{K700E} induced transcription elongation defect** (A) WT K562 cells transduced with inducible shControl (shScrambled), shDDX46, or shHTATSF1 were subjected to RT q-PCR to confirm knockdown of DDX46 (left) and HTATSF1 (right). qPCR levels normalized to housekeeping gene, GAPDH. WT+shControl levels set as 1. Data plotted are for the mean of three technical replicates. (B) Top panel: WT K562 cells electroporated with siRNA targeting control (siScr), DDX46 (siDDX46), or HTATSF1 (siHTATSF1), and SF3B1^{K700E} K562 cells electroporated with either siRNA targeting control (siScr), or HTATSF1 (siHTATSF1). Cell lysates subjected to RT q-PCR to confirm knockdown of DDX46 and HTATSF1 in WT, and HTATSF1 in SF3B1^{K700E}. qPCR levels normalized to housekeeping gene, GAPDH. siScr levels set as 1. Data plotted are for the mean of two technical replicates (for DDX46) and 3 technical replicates (for HTATSF1). Bottom panel: Western blot analysis confirming knockdown of HTATSF1 (in siHTATSF1-electroporated cells) and DDX46 (in siDDX46-electroporated cells)

with antibodies against DDX46, HTATSF1, and GAPDH. (C) Meta plots showing the log_2 fold change signal of RNAPII Ser2P ChIP-seq IP data (normalized to input) across the annotated gene transcripts genes (n = 6694 genes) SF3B1^{K700E} K562 and WT electroporated with siRNA targeting control (in WT and SF3B1^{K700E}), DDX46 (in WT), and HTATSF1 (in WT and SF3B1^{K700E}). Cells induced with doxycycline (day 0) and electroporated with siRNA on days 3 and 4, followed by ChIP on day 5. (D) Top panel: SF3B1^{K700E} K562 cells transduced with doxycycline-inducible overexpression lentiviral constructs, expressing either for empty vector (EV), N-terminal HA-tagged DDX46 (DDX46-OE), or N-terminal HA-tagged HTATSF1 (HTATSF1-OE). Cell lysates subjected to RT q-PCR to confirm overexpression of DDX46 and HTATSF1 in SF3B1^{K700E}. qPCR levels normalized to housekeeping gene, GAPDH. EV levels set as 1. Data plotted are for the mean of two technical replicates based on two independent primer sets. Bottom panel: Western blot analysis confirming overexpression of DDX46 (left) and HTATSF1 (right) in SF3B1^{K700E} +DDX46-OE and SF3B1^{K700E} + HTATSF1-OE cells, respectively, using with antibodies against HA-tag and GAPDH. (E) Gene tracks examples of CDK9 occupancy in WT and SF3B1^{K700E} cells for the indicated genes. The x-axis indicates the chromosome position, and the y-axis represents normalized read density in RPKM. (F) Quantification of readthrough phenotypes in WT and SF3B1^{K700E} cells in TT-TL seq. Distribution of read-through levels (log2 ratio of reads in read-through region vs. gene body) for top 1000 expressed genes. (G) Long-read coverage in the region downstream of PASs is shown for long-reads, in WT and SF3B1^{K700E}. Coverage is normalized to the position 100 nt upstream of each PAS (H) Top 15 biological processes of gene ontology (GO) terms sorted by Pvalue < 0.01. Counts of genes and P-value related to each BP term are shown. The y-axis represents BP terms, and the x-axis shows counts of genes were annotated to the BP terms. GO enrichment analysis performed on statistically significant upregulated genes, based on 4-day RNA-seq data, in K562 SF3B1^{K700E} (vs WT).

Figure 4: **Effect of SF3B1^{K700E} on R-loops, DNA replication dynamics and genome wide R-loop accumulation**. (A) Quantification of nuclear S9.6 intensity treated as in **C**. Data are plotted as mean of the medians + sem. ($n = 3$ individual experiments; two-tailed, unpaired Student's *t*-test; ***$P = 0.0009$, *$P = 0.016$, $^{ns}P > 0.05$). Scale bar, 2 µm. a.u: arbitrary units. (B) Representative images and quantification of RNAPII NTD + PCNA PLA foci in cells treated as in **C**. Data plotted as individual dots (each dot representing a PLA focus) (Mann-Whitney U test, two-tailed). ***P value < 0.0001, ^{ns}P value >0.05. WT cells stained with single antibody probes (PCNA only and RNAPII NTD only) included as negative controls (C) Epistatic study between SF3B1^{K700E} and known R-loop preventing factors. Left Panel: Quantification of RNAPII N-terminal domain + PCNA PLA in SF3B1^{K700E} cells, transfected with scrambled siRNA (siC), and either siRNA against Senataxin (siSETX), siRNA against FANCD2 (siFANCD2), or siRNA against UAP56 (siUAP56). siRNA transfection was performed 72 after doxycycline induction (and repeated at 96 hours), and analysis performed at 120 hours post-doxycycline induction. Data presented as box and whisker plot (n = 3 separate experiments). Scale bar, 17 µm. (Mann-Whitney U test, one-tailed, $*P_{siC\ vs\ siSenataxin} = 0.03$, $^{n.s}P_{siC\ vs\ siFANCD2} > 0.05$, $^{n.s}P_{siC\ vs}$

$_{siUAP56}$>0.05). Right panel: Quantification of average gamma-H2AX nuclear intensity in SFB1^{K700E} cells, treated as in **Left Panel**. Data are plotted as the mean of the medians + sem. ($n = 3$ individual experiments); one-tailed, unpaired Student's *t*-test; Scale bar, 17 µm. (*$P_{siC\ vs\ siSETX}$ =0.025, $^{ns}P_{siC\ vs\ siFANCD2}$>0.05, $^{ns}P_{siC\ vs\ siUAP56}$>0.05). (D) Quantification of flow cytometry cell cycle plots in WT, WT + RNH1-OE, SF3B1^{K700E}, SF3B1^{K700E} + RNH1-OE cells at 4 days and 8 days following doxycycline induction. The Watson pragmatic fitting algorithm was used for cell cycle analysis. Data presented as mean ± SEM ($n = 3$ separate experiments). Significance assessed using ANOVA followed by Sidak's test: (%$S_{4days:WT\ vs\ SF3B1^{K700E}}$ Adjusted *P*n.s; %$G1_{8days:WT\ vs\ SF3B1^{K700E}}$***Adjusted $P < 0.0001$; %$S_{8days:WT\ vs\ SF3B1^{K700E}}$***Adjusted $P < 0.0001$; %$S_{8days:SF3B1^{K700E}\ vs\ SF3B1^{K700E}\ +\ RNH1\text{-}OE}$***Adjusted $P < 0.0001$). Significance panels for S-phase and G1-phase comparisons are shown above and below the bar plots, respectively. (E) Representative images and quantification of fork velocity performed in WT, SF3B1^{K700E}, and SF3B1^{K700E} + RNH1-OE cells at 6 days post-doxycycline induction. Significance assessed using one-tailed Mann-Whitney U test (WT vs SF3B1^{K700E}***P value < 0.0001; SF3B1^{K700E} vs SF3B1^{K700E} + RNH1-OE***P value < 0.0001; WT vs SF3B1^{K700E} + RNH1-OE ^{ns}P value >0.05). (F) Representative images and quantification of nuclear EdU signal intensity and percentage of cells incorporating EdU (EdU^{+}) in WT, SFB31^{K700E}, and SF3B1^{K700E} + RNH1-OE. Scale bar, 17 µm. Bottom left panel: Percent EdU positive cell data are plotted as mean of the medians + sem. ($n = 3$ individual experiments; two-tailed, unpaired Student's *t*-test; **$P = 0.0028$, *P =0.019, ^{ns}P > 0.05). Bottom right panel: Signal EdU intensity data are plotted as a box plot ($n = 3$ individual experiments). The center line indicates a median value; the boxes and whiskers indicate 25th to 75th and 5th to 95th percentiles, respectively (one-tailed Mann–Whitney *U*-test. **$P = 0.0037$, *P =0.039,^{ns}P > 0.05). (G) Metaplot of the distribution of S9.6 signals (DRIP-seq mean coverage) along 9870 high expressing human genes and flanking regions (±10 kb) in control (blue) and SF3B1^{K700E} (orange) K562 cells. Solid lines represent the averaged signal, and the shaded area represent the upper and lower density limits of the individual replicates in each (n =2, WT; n =2, SF3B1^{K700E}). Read densities are calculated as reads per kilobase per million sequenced reads (RPKM). Differences in signal intensity at TTS ±2 kb were calculated with the Wilcoxon rank-sum test with continuity correction. Abbreviations used: TSS-transcription start site, TES-transcription end site, kb-kilobase pair. (H) Metaplot of the distribution of S9.6 signals (DRIP-seq mean coverage) along 5722 high expressing human convergent gene/gene pairs and flanking regions (±10 kb) in control (blue) and SF3B1^{K700E} (orange) K562 cells. Solid lines represent the averaged signal, and the shaded area represent the upper and lower signal limits of the individual replicates in each (n =2, WT; n =2, SF3B1^{K700E}). Read densities are calculated as RPKM. Differences in signal intensity at TTS ±2 kb were calculated with the Wilcoxon rank-sum test with continuity correction. (I) Quantification of head-on (HO; blue bar) vs co-directional (HO; orange bar) Transcription-replication conflicts intersecting with the SF3B1^{K700E} specific R-loop peaks. Absolute numbers are plotted. (J) DRIP-seq average coverage at SF3B1^{K700E} specific R-loop peaks reached by the replication fork in CD or HO orientation during DNA replication. Data plotted as box-whisker plot. (Mann-Whitney U test, two-tailed, ***P value < 0.0001).

Supplemental Figure 4: **Effect of SF3B1^{K700E} on R-loops, DNA replication dynamics and genome wide R-loop accumulation** (A) WT, SF3B1^{K700E}, SF3B1^{K700E} + RNaseH1 (RNH1) overexpression (OE) cells were induced with doxycycline for 4 days (Day 0 at 96 hours post-induction time point), and their cell growth plotted over a course of 14 days ($n = 3$, error bars indicate SEM). Significance assessed using two-tailed unpaired *t* test (***$P_{7\ days(SF3B1^{K700E}\ vs\ SF3B1^{K700E}\ +\ RNH1\text{-}\ OE)} < 0.0001$, **$P_{7\ days(WT\ vs\ SF3B1^{K700E}\ with\ RNH1\text{-}OE)} = 0.0015$, ***$P_{14\ days(SF3B1^{K700E}\ vs\ SF3B1^{K700E}\ +\ RNH1\text{-}OE)} < 0.0001$, **$P_{14\ days(WT\ vs\ SF3B1^{K700E+RNH1\text{-}OE})} = 0.0018$). (B) Nucleic acid extracted from WT (with and without RNaseH1 overexpression) and SF3B1^{K700E} (with and without RNaseH1 overexpression) cells were pre-treated with RNase III (to eliminate dsRNA) and the amount of RNA:DNA hybrids was analyzed using S9.6 antibody. Nucleic acids quantified using Nanodrop™ and loaded at 600 and 300 ng in each replicate. Additionally, 600 ng of RNaseH1 pre-treated nucleic acid extracts included for each replicate, serving as a negative control. Methylene blue staining of the nucleic acid extracts served as loading control. (C) Fluorescence microscopy images and quantification of neutral comet single cell gel electrophoresis assay in treated as in **B**. Data are plotted as the mean of the medians + sem. ($n = 3$ individual experiments; two-tailed, unpaired Student's *t*-test; ***$P < 0.0005$, $^{ns}P > 0.05$). Scale bar, 17 µm. (D) Top panel: Lysates from WT, SF3B1^{K700E}, and SF3B1^{K700E} with RNH1 overexpression cells were analyzed by western blotting (WB) for γH2AX, 6 days post induction. H2AX served as loading control. Lower panel: Relative γH2AX band intensities were quantified ($n = 3$). Significance assessed using two-tailed unpaired *t* test (*$P_{\gamma H2AX(SF3B1^{K700E}\ vs\ WT)} = 0.02$, (*$P_{\gamma H2AX(SF3B1^{K700E}\ vs\ SF3B1^{K700E}\ with\ RNH1\text{-}\ OE)} = 0.03$, $^{ns}P_{\gamma H2AX(WT\ vs\ SF3B1^{K700E}\ with\ RNH1\ \text{-}OE)} = 0.57$;error bars indicate SEM). (E) IF images and quantification of BrdU and γH2AX immunofluorescence in SF3B1^{K700E} cells shown as mean ± sem ($n = 3$ separate experiments). a.u: arbitrary units. Significance assessed using one-tailed unpaired *t* test (***$P < 0.0001$). Scale bar = 15 µm. (F) IF images and quantification of pRPA and γH2AX immunofluorescence in SF3B1^{K700E} cells at 6 days post induction, shown as mean ± SEM ($n = 3$ separate experiments). *$P = 0.02$. a.u: arbitrary units. Significance assessed using one-tailed unpaired *t* test. Scale bar = 15 µm. (G) IF images and quantification of S9.6 and γH2AX IF in SF3B1^{K700E} cells at 6 days post induction, shown as shown as mean ± sem ($n = 3$ separate experiments). a.u: arbitrary units. Significance assessed using one-tailed unpaired *t* test **$P = 0.0011$. (H) Western Blot-based confirmation of depletion of Senataxin, FANCD2, and UAP56, by siRNA knockdown experiments. siRNAs and antibodies used are indicated. (I) Epistatic study between SF3B1^{WT} and known R-loop preventing factors in K562 cells. Left Panel: Representative images and quantification of RNAPII NTD +PCNA PLA in WT cells, transfected with scrambled siRNA (siC), and either siRNA against Senataxin (siSETX), siRNA against FANCD2 (siFANCD2), or siRNA against UAP56 (siUAP56). siRNA transfection by electroporation was performed 72 and 96 hours after doxycycline induction, and analysis performed at 120 hours post-doxycycline induction. Data presented as box and whisker plot (n = 3 separate experiments). Scale bar, 17 µm. (Mann-Whitney U test, one-tailed, ***$P_{siC\ vs\ siSETX} < 0.0001$, ***$P_{siC\ vs\ siFANCD2} < 0.0001$, ***$P_{siC\ vs\ siUAP56} < 0.0001$). Right Panel: Representative images and quantification of average gamma-H2AX nuclear intensity in WT cells

treated as in **Left Panel.** Data are plotted as the mean of the medians + sem. ($n = 3$ individual experiments; one-tailed, unpaired Student's *t*-test; Scale bar, 17 µm. (**$P_{siC\ vs\ siSenataxin} < 0.0012$, ***$P_{siC\ vs\ siFANCD2} < 0.0057$, ***$P_{siC\ vs\ siUAP56} < 0.0001$). (J) SF3B1^{K700E} K562 cells transduced with E. coli RNaseH1-expressing retroviral vectors were subjected to RT q-PCR to confirm bacterial RNaseH1 overexpression. qPCR levels normalized to housekeeping gene, GAPDH. Human RNaseH1 levels are set as 1. Data plotted are for the mean of two biological replicates. (K) Representative images of flow cytometry cell cycle plots in WT, WT + RNH1-OE, SF3B1^{K700E}, SF3B1^{K700E} + RNH1-OE cells at 4 days and 8 days following doxycycline induction. Cell counts plotted on Y axis and DAPI content plotted on X axis. (L) IF images and quantification of EdU positive SF3B1^{K700E} cells, 6 days post induction, that are stained for γH2AX, CC3 and EdU. Data comparing gamma-H2AX intensity in the EdU +ve versus EdU -ve cells are plotted as mean of the medians + sem. ($n = 3$ individual experiments; two-tailed, unpaired Student's *t*-test; **$P = 0.0011$). Scale bar, 15 µm. a.u: arbitrary units. (M) Schematic of triptolide treatment experiment in G1/S synchronized SF3B1^{K700E} cells. Bottom left panel: Quantification of ethynyl uridine (EU) and gamma-H2AX nuclear intensities in SF3B1^{K700E} cells treated with either Triptolide and DMSO or DMSO alone. IF images of EU and γH2AX fluorescence. Data are plotted as mean of the medians + sem. ($n = 3$ individual experiments; two-tailed, unpaired Student's *t*-test; ***$P < 0.0001$). Scale bar, 2 µm. a.u: arbitrary units. Bottom right panel: Quantification of RNAPII + PCNA PLA foci. Foci number per cell are presented as a box-whisker plot ($n = 3$; two-tailed Mann–Whitney *U*-test). The median values are indicated by black lines. Scale bar, 2 µm. ***$P < 0.0001$. (N) Left: Percentage of new origins of replication (CldU only tracks) performed in WT, SF3B1^{K700E}, and SF3B1^{K700E} + RNH1-OE cells at 6 days post-doxycycline induction. Right panel: Quantification of DNA fiber track length (tracks with both CldU and IdU labeling) performed in WT, SF3B1^{K700E}, and SF3B1^{K700E} + RNH1-OE cells at 6 days post-doxycycline induction. Significance assessed using one-tailed Mann-Whitney U test (WT vs SF3B1^{K700E}***P value < 0.0001; SF3B1^{K700E} vs SF3B1^{K700E} + RNH1-OE***P value < 0.0001; WT vs SF3B1^{K700E} + RNH1-OE ^{ns}P value >0.05). (O) Scatter plots depicting correlation between DRIP-seq samples in WT and SF3B1^{K700E} cells. Values in boxes represent Pearson's correlation coefficients between corresponding samples ($n = 2$ biological replicates). (WT, r=0.99; SF3B1^{K700E}, r=0.99). Values expressed as Log2(gene-expression + 1) (P) RT–qPCR analyses of DRIP IP/Input of EGR1, snRPN, RPL13, and TFPT genes in WT and SF3B1^{K700E} cells, 4 days post doxycycline induction. EGR1 and snRPN represent low R-loop signal genes while TFPT and RPL13 denote high R-loop signal genes. snRPN (IP/input) was set as 1. WT (n=2, WT-R1 and WT-R2) and SF3B1^{K700E} (n=2, SF3B1^{K700E}-R1 and SF3B1^{K700E}-R2). Data plotted are for the mean of two technical replicates. (Q) Volcano plot of DRIP-seq differential analysis between SFB1^{K700E} versus WT K562 cells (n = 2 biological replicates, each). Dots represent peaks. Abbreviations used: FDR-false discovery rate, L2FC-log2 fold change. (R) Genome annotation of R-loop gain peaks in WT and SF3B1^{K700E} K562 cells using HOMER software. Promoter: -100bp to +300bp of transcription start site, Gene body: TSS + 300bp to transcription end site (TES), Terminator: TES to +300bp downstream. (S) Categorization of gene pairs based on orientation of transcription between neighboring genes (Genes 1 and 2): (a) Convergent gene-pairs characterized by

converging orientation of gene transcription units, (b) and (d) Tandem gene-pairs in which neighboring genes transcribe in the same orientation, and (d) Divergent gene-pairs in which neighboring genes transcribe in diverging orientations. Transcription termination sites (TTSs) of convergent gene-pairs represent hot-spots for head-on collisions due to the transcription associated R-loops from Gene1 blocking progression of the oppositely oriented replication fork from Gene2, and vice-versa. (T) Metaplot of the distribution of S9.6 signals (DRIP-seq mean coverage) along 5854 high expressing human non-convergent gene pairs (RPKM > 1000) and flanking regions (±5 kb) in control (blue) and SF3B1^{K700E} (orange) K562 cells. Solid lines represent the averaged signal, and the shaded area represents the upper and lower density limits of the individual replicates in each (n =2, WT; n =2, SF3B1^{K700E}). Read densities are calculated as RPKM. Differences in signal intensity at TTS ±2 kb were calculated with the Wilcoxon rank-sum test with continuity correction. (U) Metaplot of the distribution of S9.6 signals (DRIP-seq mean coverage) along high expressing human convergent gene/gene pairs and flanking regions (±5 kb), depending on the distance between their TESs, in control (blue) and SF3B1^{K700E} (orange) K562 cells. Solid lines represent the averaged signal, and the shaded area represent the upper and lower signal limits of the individual replicates in each (n =2, WT; n =2, SF3B1^{K700E}). Read densities are calculated as RPKM. Left Panel: >5kb, n=3456 genes; middle panel: 2-5kb, n=799 genes; right panel: <2kb, n=1250 genes. Abbreviations used: TSS-transcription start site, TES-transcription end site. (V) Distribution of DRIP-seq in SF3B1^{K700E} (orange) and WT (blue) K562, Okazaki fragments (OK-seq) in WT K562, and nascent transcription (GRO-seq) signals in WT K562, at two converging genes LUZP1 and KDM1A on chromosome 1 in control K562 cells. (W) Western blot analysis of WT, SF3B1^{K700E} and SF3B1^{K700E} + RNH1-OE cells, 6 days post induction, for phospho-ATM, phospho-Chk1, and phospho-Chk2 levels. ATM, Chk1 and Chk2 used as loading controls, respectively.

Figure 5: **Transcription-coupled changes in chromatin accessibility and histone marks due to SF3B1^{K700E}** (A) Heat map plots depict H3K4me3 signal in WT (left) and SF3B1^{K700E} K562 (right) at 4 days post doxycycline induction. Signals mapped from −1 kb upstream of TSS (transcription start site) to 2 kb downstream of the TES (transcription end site). Genes ranked by decreasing H3K4me3 occupancy. (B) Metagene plot showing the normalized H3K4me3 CutnRun levels over 6694 expressed genes in WT (blue) and SF3B1^{K700E} (orange). Solid lines represent the averaged signal, and the shaded area represents the upper and lower signal limits among the individual replicates (n =2, WT; n =2, SF3B1^{K700E}). IgG included as control (shown in black). Read densities are calculated as RPKM. (C) Representative genome browser track examples of H3K4me3 CutnRun-seq in K562 cells WT (blue) versus SF3B1^{K700E} (orange), at 4 days post doxycycline induction, for the indicated genes (n=2 biological samples, each). The x-axis indicates the chromosome position, and the y-axis represents normalized RPKM. (D) Metagene and heatmap plots depict H3K4me3 (left) and H3K27me3 (right) signals in WT and SF3B1^{K700E} K562 cells, at 4 days post doxycycline induction, at bivalently marked gene promoters (n=2168 genes). Signals mapped from −1 kb to +1 kb of TSS. Read

densities are calculated as RPKM. (E) Heat maps depict change in RNAPII NTD ChIP and H3K4me3 CutnRun average signal enrichments (WT – SF3B1^{K700E}) in K562 cells. Genes separated into 3 groups, based on the clusters, shown in Figure 5E. Heatmaps depict the change in RNAPII NTD ChIP-seq (left) or H3K4me3 (right) signals (WT - SF3B1^{K700E}) from −1 kb upstream of TSS (transcription start site) to 2 kb downstream of the TES (transcription end site). (F) Volcano plot of ATAC-seq differential analysis between SFB1^{K700E} versus WT K562 cells (n = 2 biological replicates, each). Dots represent peaks. Abbreviations used: FDR-false discovery rate. (G) RegioneR-based colocalization analysis of selectively downregulated ATAC peaks in SF3B1^{K700E} and DRIP-seq identified differential R-loop regions (SF3B1^{K700E} vs WT). (H) Metaplot of WT and SF3B1^{K700E} ATAC-seq signal profiling changes in the nucleosome occupancy among high expressing genes (FPKM>1000, *n* = 274 from dataset of 6694 genes). Nucleosome signal was generated using the NucleoATAC package. (I) Left: Metaplots and heatmaps of RPKM-normalized GRO-seq signal for sense transcription at promoters (left) and putative enhancers (right), ranked by decreasing occupancy. (J) Fourteen chromatin states learned jointly for WT and SF3B1^{K700E} cells by a multivariate hidden Markov model. The heatmap shows the state emissions learned on the basis of combinatorial analysis of genome-wide R-loops, RNAPII activity, ATAC and the other epigenetic modifications, where each column corresponds to a different epigenetic signature and each row corresponds to a different chromatin state. Candidate state annotations for each state are listed to the left. Blue shading indicates intensity. Putative state annotations: 1. ReprPC- repressed polycomb, 2. Quies-quiescent, 3. Rloop_only: R-loop only state without histone marks, 4. Low_signal: low signal state, 5. TxEnh- transcribing enhancers, 6. Tx 3'- transcribing gene 3'end, 7. Str_Tx-Strongly transcribing genic region, 8. Wk_Tx-Weakly transcribing genic region, 9. TxEnhA-active transcribing enhancers, 10. EnhP- poised enhancers, 11. TSSfl_Down- Region flanking downstream of TSS, 12. TSSfl-Region flanking TSS, 13. TSSA-active promoters, 14.Enh- Enhancer. (K) Relative enrichment of chromatin state transitions at promoters of downregulated genes (left panel) or upregulated genes (right panel) in SF3B1^{K700E} (vs WT), for chromatin state transition pairs involving states 10, 11, 12, and 13. Red shows enrichment, whereas blue shows depletion. For further details, see extended supplemental figure 9 data.

Supplemental figure 5: **Transcription-coupled changes in chromatin accessibility and histone marks due to SF3B1^{K700E} (**A) Scatter plots depicting correlation between the histone mark CutnRun samples in WT and SF3B1^{K700E} samples. Values in boxes represent Pearson's correlation coefficients between corresponding samples. (B) H3K4me3 CUT&RUN enrichment profiles at 5' splice sites (5'SS) of first, second, and third introns in WT and SF3B1^{K700E}. Y-axis shows the RPKM. (C) Metagene plot showing the normalized H3K27ac CUT&RUN levels over 6694 expressed genes in WT (blue) and SF3B1^{K700E} (orange) (n =1, WT; n =1, SF3B1^{K700E}). Read densities are calculated as RPKM. (D) RegioneR-based colocalization analysis of downregulated ATAC peaks with upregulated R-loop regions (SF3B1^{K700E}/WT) at promoter and gene body regions. (E) Metaplots and heatmaps of RPKM-normalized GRO-seq signal for antisense transcription at promoters (left) and enhancers (right) ranked by decreasing occupancy. (F) Representative genome browser

track examples of GRO-seq densities and ATAC-seq in WT (blue) versus SF3B1^{K700E} (orange), for the indicated putative enhancers. The x-axis indicates the chromosome position, and the y-axis represents normalized RPKM. (G) Transition parameters for the 14-state model derived by ChromHMM for WT and SF3B1^{K700E} cells. (H) Overlap of different genomic features (Centromeres, CpG island, Putative Enhancer, Exons, Gene body, Intergenic, laminB lads, Promoter, TSS, TSS+2kb downstream, and terminator regions with chromatin state calls in WT (left) and SF3B1^{K700E} (right). The fold enrichments are calculated as the ratio between observed and expected number of genomic bins for each overlap. The color intensities are normalized within each column between its minimum value (white) and its maximum value (blue). (I) Log2 ratios between the number of total genomic bins for each chromatin state in WT and SF3B1^{K700E} cells (WT/SF3B1^{K700E}). Log2ratio plotted on Y-axis and chromatin state plotted on X-axis. A log2ratio above zero indicates a higher genomic bin count in the WT for a particular chromatin state. Conversely, a log2ratio below zero indicates a lower genomic bin count in SF3B1^{K700E} for a particular chromatin state.

Figure 6: **An epigenetic shRNA rescue screen identifying key pathways functionally important to SF3B1^{K700E}-related cell survival** (A) shRNA screen for epigenetic/chromatin modifier genes that modulate sensitivity to SF3B1^{K700E}. K562 cells (SF3B1^{K700E}, n = 3; WT, n = 1) were transduced with an shRNA lentiviral library (puro-resistant), selected in puromycin for 3 days, then induced with doxycycline (Day 0), and monitored for growth through 24 days from doxycycline induction. (B) Waterfall bubble plot shows combined SF3B1^{K700E}-selective survival effect of each gene, calculated as a weighted effect of knockdown by multiple shRNAs. SF3B1^{K700E} high-confidence candidates promoting survival (average log fold enrichment of 4 or greater [day 24 versus day 0]) are shown in red and SF3B1^{K700E} candidates inhibiting survival (with average log fold enrichment of -4 or lower [day 24 versus day 0]) are show in blue; rest are show in green. (C) Sin3 (marked as violet) and HDAC (marked as red) complex components are enriched among genes that promote cell survival in SF3B1^{K700E}. StringDB functional enrichment analysis was performed on the top 20 survival candidates identified by the shRNA screen (Log2 fold enrichment > 4 and shRNAs => 3 against a particular target). Top 5 functionally enriched cellular component pathways with FDR < 0.001 are shown. (D) WT, SF3B1^{K700E} + shScrambled, SF3B1^{K700E} + shING2, SF3B1^{K700E} + shHDAC2, SF3B1^{K700E} + shPHF21A cells were induced with doxycycline for 4 days (Day 0 at 96 hours post-induction time point), and their cell growth plotted over a course of 14 days ($n = 3$, error bars indicate SEM). Significance assessed using two-tailed unpaired t test (***$P_{\text{7 days(SF3B1}^{K700E}\text{ + shScrambled vs SF3B1}^{K700E}\text{ + shING2)}} < 0.0001$, (***$P_{\text{7 days(SF3B1}^{K700E}\text{ + shScrambled vs SF3B1}^{K700E}\text{ + shHDAC2)}} < 0.0001$, (***$P_{\text{7 days(SF3B1}^{K700E}\text{ + shScrambled vs SF3B1}^{K700E}\text{ + shPHF21A)}} < 0.0001$); ***$P_{\text{14 days(SF3B1}^{K700E}\text{ + shScrambled vs SF3B1}^{K700E}\text{ + shING2)}} < 0.0001$, (***$P_{\text{14 days(SF3B1}^{K700E}\text{ + shScrambled vs SF3B1}^{K700E}\text{ + shHDAC2)}} < 0.0001$, (***$P_{\text{14 days(SF3B1}^{K700E}\text{ + shScrambled vs SF3B1}^{K700E}\text{ + shPHF21A)}} < 0.0001$). (E) Representative images and quantification of RNAPII NTD + PCNA PLA foci in SF3B1^{K700E} + shScrambled (shScr), SF3B1^{K700E} +

shHDAC2-1, SF3B1^{K700E} + shING2-1 , SF3B1^{K700E} + shPHF21A at 4 days post induction. Data plotted as individual dots (each dot representing a PLA focus) (Mann-Whitney U test, two-tailed). ***P value < 0.0001. Scale bar 15 um. (F) Metagene plot analysis showing the $\log_2$ fold change (L2FC) of anti-RNAPII Ser2P ChIP-seq data normalized to input across the annotated gene transcripts genes (n = 6694 genes) in WT, SF3B1^{K700E} + shScr, SF3B1^{K700E} + shHDAC2, SF3B1^{K700E} + shING2, SF3B1^{K700E} + shPHF21A K562 cells. Abbreviations used: TSS-transcription start stie, TES-transcription end site. (G) Metagene plot analysis showing the RPKM normalized H3K4me3 CutnRun levels over 6694 expressed genes in WT, SF3B1^{K700E} + shScr, SF3B1^{K700E} + shHDAC2, SF3B1^{K700E} + shING2, SF3B1^{K700E} + shPHF21A K562 cells. Read densities are calculated as RPKM. (H) Total RNA from the indicated cell lines was extracted at 5 days post doxycycline induction, followed by cDNA RT-PCR of the cryptic 3′SSs (black arrowheads) and canonical 3’SSs (blue arrowheads) of six select genes (*TTI1, PPP245A, DYNLL1, ABCB7, ENOSF1, MAP3K7* from top to bottom).

Supplemental Figure6: **An epigenetic shRNA rescue screen identifying key functional pathways important to SF3B1^{K700E}-related cell survival** (A) Plot correlation heatmap of SF3B1^{K700E} (n=3 per genotype; labeled as MUT.1, MUT.2, MUT.3) at different time points of gDNA analysis (days 0, 8, 17, 24). (B) GO term enrichment analysis of top 20 survival candidates. Enriched GO terms (FDR <0.05) grouped by biologic process function. X axis represents the number of genes representing a particular GO term function. (C) Western blot confirmation of shRNA mediated knockdown of protein levels. Cells transduced with shScr or two independent shRNAs knocking down (KD) HDAC2 (shHDAC2-1 and shHDAC2-2), ING2 (shING2-1 and shING2-2), and PHF21A (shPHF21A-1 and shPHF21A-2) were subjected to western blotting with respective antibodies. Of note, subsequent analysis was restricted to the stronger shPHF21A-2 clone based on effectiveness in depleting PHF21A protein levels. GAPDH used as loading control for testing HDAC2 and PHF21A KDs, and beta-actin used as loading control for testing ING2 KD. (D) Representative images and quantification of nuclear S9.6 intensity in SF3B1^{K700E} + shScrambled (shScr), SF3B1^{K700E} + shHDAC2-1, SF3B1^{K700E} + shING2-1 , SF3B1 ^{K700E} + shPHF21A at 6 days post induction. Data are plotted as mean of the medians + sem. ($n = 3$ individual experiments; two-tailed, unpaired Student’s *t*-test; $^{**}P_{(SF3B1^{K700E} + shHDAC2 \text{ vs } SF3B1^{K700E} + shScrambled}$ = 0.0031; $^{**}P_{(SF3B1^{K700E} + shING2 \text{ vs } SF3B1^{K700E} + shScrambled}$ = 0.004; $^{**}P_{(SF3B1^{K700E} + shPHF21A \text{ vs } SF3B1^{K700E} + shScrambled}$ = 0.0013). Scale bar,15 µm. a.u: arbitrary units. (E) Fluorescence microscopy images of neutral comet single cell gel electrophoresis assay in SF3B1^{K700E} + shScrambled (shScr), SF3B1^{K700E} + shHDAC2, SF3B1^{K700E} + shING2 , SF3B1^{K700E} + shPHF21A cells at 6 days post induction. Data are plotted as mean of the medians + sem. ($n = 3$ individual experiments; two-tailed, unpaired Student’s *t*-test; $^{**}P_{(SF3B1^{K700E} + shScr \text{ vs } SF3B1^{K700E} + shING2)}$= 0.001; $^{***}P_{(SF3B1^{K700E} + shScr \text{ vs } SF3B1^{K700E} + shHDAC2)}$<0.001; $^{**}P_{(SF3B1^{K700E} + shScr \text{ vs } SF3B1^{K700E} + shPHF21A)}$= 0.004). Scale bar, 17 µm. (F) IF images and quantification of γH2AX fluorescence in SF3B1^{K700E} + shScrambled (shScr), SF3B1^{K700E} + shHDAC2-1, SF3B1^{K700E} + shING2-1 , SF3B1^{K700E} + shPHF21A at 5 days post induction. Data are plotted as mean of the medians + sem. ($n = 3$ individual

experiments; two-tailed, unpaired Student's *t*-test; $^{**}P_{(SF3B1^{K700E} + shHDAC2 \text{ vs } SF3B1^{K700E} + shScrambled}$ = 0.0038; $^{**}P_{(SF3B1^{K700E} + shING2 \text{ vs } SF3B1^{K700E} + shScrambled}$ = 0.006; $^{**}P_{(SF3B1^{K700E} + shPHF21A \text{ vs } SF3B1^{K700E} + shScrambled}$ = 0.005). Scale bar,15 µm. a.u: arbitrary units. (G) Representative images and quantification of flow cytometry cell cycle plots in $SF3B1^{K700E}$ + shScrambled (shScr), $SF3B1^{K700E}$+ shHDAC2, $SF3B1^{K700E}$ + shING2 , $SF3B1^{K700E}$ + shPHF21A cells at 4 days and 8 days following doxycycline induction. The Watson pragmatic fitting algorithm was used for cell cycle analysis. Data presented as mean ± sem (n=3 separate experiments). Significance assessed using ANOVA followed by Sidak'stest: ($\%G2/M_{4days:SF3B1^{K700E} + shScr \text{ vs } SF3B1^{K700E} + shPHF21A}$ =0.04; $\%G1_{8days:SF3B1^{K700E}+ shScr \text{ vs } SF3B1^{K700E}+ shING2}$****Adjusted $P < 0.0001$; $\%G1_{8days:SF3B1^{K700E}+ shScr \text{ vs } SF3B1^{K700E}+ shHDAC2}$****Adjusted $P < 0.0001$; $\%G1_{8days:SF3B1^{K700E}+ shScr \text{ vs } SF3B1^{K700E}+ shPHF21A}$****Adjusted $P < 0.0001$; $\%S_{8days:SF3B1^{K700E}+ shScr \text{ vs } SF3B1^{K700E} + shING2}$****Adjusted $P < 0.0001$; $\%S_{8days:SF3B1^{K700E}+ shScr \text{ vs } SF3B1^{K700E}+ shHDAC2}$****Adjusted $P < 0.0001$; $\%S_{8days:SF3B1^{K700E}+ shScr \text{ vs } SF3B1^{K700E}+ shPHF21A}$****Adjusted $P < 0.0001$) (H) Quantification of percentage of cells incorporating EdU (EdU^+) in $SF3B1^{K700E}$ + shScrambled (shScr), $SF3B1^{K700E}$ + shHDAC2, $SF3B1^{K700E}$ + shING2 , $SF3B1^{K700E}$ + shPHF21A cells at 6 days following doxycycline induction. Percent EdU positive cell data are plotted as mean of the medians + sem. ($n = 3$ individual experiments; two-tailed, unpaired Student's *t*-test; $^{**}P_{(SF3B1^{K700E} + shScr \text{ vs } SF3B1^{K700E} + shING2)}$= 0.0011; $^{**}P_{(SF3B1^{K700E} + shScr \text{ vs } SF3B1^{K700E} + shHDAC2)}$= 0.004; $^{**}P_{(SF3B1^{K700E} + shScr \text{ vs } SF3B1^{K700E} + shPHF21A)}$= 0.0062). Scale bar, 15 µm. (I) Quantification of fork velocity performed in $SF3B1^{K700E}$ + shScrambled (Tangerine), $SF3B1^{K700E}$ + shING2 (olive green), $SF3B1^{K700E}$ + shHDAC2 (blue violet) , $SF3B1^{K700E}$ + shPHF21A (neon green) cells at 6 days post induction. Data are shown as a box plot ($n = 2$). Significance assessed using one-tailed Mann-Whitney U test ($SF3B1^{K700E}$ + shScr vs $SF3B1^{K700E}$+ shING2,$^{***}P$ value < 0.001; $SF3B1^{K700E}$ + shScr vs $SF3B1^{K700E}$ + shHDAC2,$^{***}P$ value < 0.001; $SF3B1^{K700E}$ + shScr vs $SF3B1^{K700E}$ + shPHF21A,$^{**}P$ value =0.001). (J) Read coverage in the introns undergoing cryptic 3'SS splicing [determined from analysis of patient sample data] (left) versus unselected introns (right). Coverage is normalized to the position 100 nt upstream of each 5'SS. Orange line indicates $SF3B1^{K700E}$ and the blue line indicates WT. (K) The cumulative distribution function plot of the TR distribution, calculated as an average of two replicates) in WT and $SF3B1^{K700E}$ cells depending on genes with introns undergoing cryptic 3'SS splicing (determined from analysis of MDS patient sample data) versus unselected genes. Significance determined using two sample K.S. test.

Figure7: **Targeting H3K4me pathway in SF3B1-mutant MDS.** (A) Workflow for isolation/characterization of SF3B1- mutant MDS patient marrow aspirate derived CD34+ cells. (B) Metagene plot analysis showing RPKM normalized CUT&RUN levels of RNAPII Ser2P over 6694 expressed genes in WT (blue) and SF3B1-mutant (orange) MDS CD34+ cells. Solid lines represent the averaged signal, and the upper and lower borders of the shaded area represent the highest and lowest density ranges among the replicates (n =5 experiments, WT; n =7 experiments, SF3B1-mutant MDS). Read densities are calculated as reads per RPKM. (C) Representative genome browser track examples of Ser2P RNAPII CUT&RUN in CD34+ cells: WT (blue) versus $SF3B1^{K700E}$ (orange), for the indicated genes. Corresponding IgG tracks for the indicated genes in

black. The x-axis indicates the chromosome position, and the y-axis represents normalized RPKM. (D) Left: Metagene plot showing RPKM normalized CUT&RUN levels of H3K4me3 over 6694 expressed genes in WT (blue) and SF3B1-mutant (orange) MDS CD34+ cells. Solid lines represent the averaged signal, and the upper and lower borders of the shaded area represent the highest and lowest density ranges among the replicates, combined (n =4 experiments, WT; n =5 experiments, SF3B1^{K700E} MDS). Read densities are calculated as RPKM. Right: Box whisker plots of H3K4me3 promoter density (RPKM) in WT (n=4) and SF3B1- mutant MDS (n=5) CD34+ cells. (student t test, two-tailed, $^{**}P$=0.0072). (E) Representative genome browser track examples of H3K4me3 CUT&RUN in CD34+ cells WT versus SF3B1-mutant, for the indicated genes. The x-axis indicates the chromosome position, and the y-axis represents normalized read density in reads per kilobase pair per million (RPKM). F) Metaplot of WT and SF3B1-mutant ATAC-seq signal profiling changes in the nucleosome occupancy among 6694 expressing genes in WT (n=3 experiments) and SF3B1-mutant (n = 6 experiments) CD34+ cells. Nucleosome signal was generated using the NucleoATAC package. (G) Methylcellulose colony forming unit assay of mononuclear cells isolated from healthy volunteer and MDS patient derived peripheral blood/bone marrow aspirates. Isolated cells were treated with either 0.05% DMSO or 5uM OICR-9429 (in final 0.05% DMSO concentration) on day 0 after isolation and incubated in methylcellulose medium for 14 days before colony counting and imaging. Left panel: representative images of WT (top) and SF3B1-mutant (bottom) exposed to either DMSO (control) or 5uM OICR-9429. Cell counting analysis performed manually using stem cell grid. Right panel: Colony formation analysis in WT (n = 4 independent experiments) and SF3B1-mutant (n = 8 independent experiments) treated with OICR-9429 (relative to DMSO controls). $^{***}P$= 0.0003 (two tailed student's t test). (H) Methylcellulose colony forming unit assay of mononuclear cells isolated from MDS patient derived peripheral blood/bone marrow aspirates. Isolated cells were transduced on day 0 with GFP-expressing shRNA lentivirus and then incubated in methylcellulose medium for 14 days before colony counting and imaging. Representative images of sh scrambled (top left), sh HDAC2 (top right) and sh ING2 (bottom left). Cell counting analysis performed manually using stem cell grid. Right lower panel: Colony formation analysis in sh scrambled, shHDAC2, shING2 (all relative to sh scrambled). $^{*}P_{shHDAC2vsScrambled} = 0.01$, $^{*}P_{shING2vsScrambled}$ = 0.05 (one-tailed Student's t test). (I) Experimental scheme for murine experiments. Adult (6-8 week) *R26*-CreER$^{T2\ KI/+}$ *Sf3b1*$^{+/K700E}$ mice and age-matched *R26*-CreER$^{T2\ KI/+}$ *Sf3b1*$^{+/+}$ controls were injected with tamoxifen (in corn oil) daily for 5 days. Peripheral blood was collected 2 weeks later and cDNA sequencing of Sf3b1K700 locus region performed to confirm mutant allele expression. LSK population (cKit+Sca-1+Lin- cells) were isolated from marrow cells by FACS and divided for CFU or ATAC-seq. (J) Methylcellulose colony forming unit assay of LSK cells isolated from murine Sf3b1^{K700E} bone marrow aspirates. Isolated cells were transduced on day 0 with lentiviruses expressing shRNA and then incubated in methylcellulose medium for 14 days before colony counting and imaging. Representative images of sh scrambled, sh Hdac2, shIng2, and shWdr5 are shown. Cell counting analysis performed manually using stem cell grid. Right lower panel: Colony formation analysis in shScrambled, shHdac2, shIng2, shWdr5 (all relative to shScrambled). $^{**}P_{shWdr5vsshScrambled} < 0.01$(one-tailed

Student's t test). (K) Metaplot of WT and Sf3b1^{K700E} ATAC-seq signal profiling changes in the nucleosome occupancy among 1024 high-expressing genes (TPM>1000) in WT and Sf3b1^{K700E} cKit+Sca-1+Lin- murine cells. Nucleosome signal was generated using the NucleoATAC package.

Supplemental Figure 7: **Targeting H3K4me pathway in SF3B1-mutant MDS.** (A) Gating strategy used throughout the experiments for flow cytometric analysis of CD34 surface expression markers on mononuclear cells isolated from bone marrow aspirates. Mononuclear cells were stained with DAPI and acquired in the flow cytometer. Events were gated based on forward (FSC) and side (SSC) scatter characteristics and then gated on live cells (DAPI -). Live cells were then gated on CD34 surface marker expression. Top panel: representative example of gating strategy in CD34 unstained negative control cell population; Bottom panel: representative example of gating strategy in one of the SF3B1-mutant MDS marrow aspirate derived mononuclear cells. (B) Quantification of RNAPII NTD + PCNA foci (by PLA assay) in WT and SF3B1-mutant MDS derived cells. Foci number per cell are presented as a scatter plot ($n = 3$ independent experiments each; two-tailed Mann–Whitney *U*-test). The median values are indicated by black lines. Scale bar, 2 µm. ***$P < 0.0001$. (C) Quantification of gamma-H2AX intensity in WT and SF3B1-mutant MDS derived cells. Gamma-H2AX intensity presented as a bar plot. Mean + sem ($n = 3$ individual experiments). Scale bar, 2 µm. ***$P < 0.0001$. (D) Representative example showing chromatograms of PCR amplicons around the SF3B1 R625 locus. Patient-derived SF3B1^{R625C} MDS CD34^{+} cells were isolated and grown in methylcellulose medium for 14 days after being transduced with lentiviral constructs targeting non-specific control (shScrambled), HDAC2 (shHDAC2), ING2 (shING2). DNA was isolated from the cell colonies at 14 days for the respective conditions and PCR performed on the extracted DNA to amplify the region around the SF3B1 R625 locus. Highlighted in blue is the R625 locus, showing the presence of TGT mutation (R→C). (E) Gating strategy used throughout the experiments for flow cytometric analysis of C11b and CD71 surface expression markers on nucleated cells isolated from SF3B1-mutant MDS derived colonies grown in methylcellulose medium for 14 days. Mononuclear cells were stained with DAPI and acquired in the flow cytometer. Events were first gated based on forward (FSC) and side (SSC) scatter characteristics and then gated on live cells (DAPI -). Live cells were then gated on the FITC channel for GFP-expressing cells. Flow cytometric evaluation on the GFP+ cells was performed to evaluate for C11b (APC) and CD71 (PE-Cy7) surface expression. Panels from top to bottom in the following order: unstained control, SF3B1-mutant MDS derived cells transduced with GFP-expressing shControl, shHDAC2, and shING2 lentiviral vectors, respectively. (F) Percentage of CD11b and CD71 expressing cells (among live, mononuclear cells) in shControl, shHDAC2, and shING2 settings. Data presented as the mean of 3 individual patient samples. Each datapoint represents a patient sample. Statistical comparisons deferred due to high variability across biological replicates. (G) Gating strategy used throughout the experiments for flow cytometric analysis and isolation of cKit+Sca-1+Lin- cells isolated from bone marrow aspirates. Mononuclear cells were stained with DAPI and acquired in the flow cytometer. Events were

gated based on forward (FSC) and side (SSC) scatter characteristics and then gated on live cells (DAPI -). Live cells were then gated on Sca-1 surface marker expression (FITC). Sca-1 expressing cells were then gated on cKit (PE-Cy7) and Lineage cocktail (APC) marker expressions. Top panel: representative example of gating strategy in unstained negative control cell population; Bottom panel: representative example of gating strategy in one of the marrow mononuclear cells to isolate cKit+Sca-1+Lin- cells. (H) Methylcellulose colony forming unit assay of cKit+Sca-1+Lin- cells isolated from murine Sf3b1 and WT bone marrow aspirates. Isolated cells were transduced on day 0 with lentiviruses expressing shRNA and then incubated in methylcellulose medium for 14 days before colony counting and imaging. Representative images of sh scrambled, sh Hdac2, shIng2, and shWdr5 are shown. Cell counting analysis performed manually using stem cell grid. Right lower panel: Colony formation analysis in shScrambled, shHdac2, shIng2, shWdr5 (all relative to shScrambled). (I) Gating strategy used throughout the experiments for flow cytometric analysis of C11b and CD71 surface expression markers on nucleated cells isolated from murine Sf3b1^{K700E} cKit+Sca-1+Lin- cell-derived colonies grown in methylcellulose medium for 14 days. Mononuclear cells were stained with DAPI and acquired in the flow cytometer. Events were first gated on live cells (DAPI-) and then based on forward (FSC) and side (SSC) scatter characteristics. An increase in the percentage of events was noted in the 'monocyte' gate (low SSC, high FSC), for the shHdac2 and shIng2 settings. Flow cytometric evaluation on the cells, evaluating for C11b (PE-Cy7) and CD71 (APC) surface expression, showed that the events in this gate indeed corresponded to CD11b+ monocytes. Panels from top to bottom in the following order: unstained control, Sf3b1^{K700E} cKit+Sca-1+Lin- cells transduced with shScrambled, shHdac2, shIng2, and shWdr5 expressing lentiviral vectors, respectively.

Supplemental FigS8: Unsupervised clustering of alternative splicing events across 172 MDS/AML patients, and 45 normal samples from 3 individual datasets[33 35 34]. Alternative splicing events were first determined by rMATS package individually for each sample. Events which were present at PSI (percentage spliced in or fraction of mRNA that corresponds to the novel isoform) value of 0.1 or higher in at least 10% of the total sample were selected (total 137,435 events) and unsupervised clustering performed by ComplexHeatmap R package. Mutational data for SF3B1, U2AF1, SRSF2 and ZRSR2 are shown on the top panel. Alternative splicing event type are show on the right-side column (A3SS = alternative 3' Splice Site, A5SS = alternative 5' splice site, MXE = mutually exclusive exon, SE = skipped exon, RI = retained intron)

REFERENCES:

1. Reimer, K.A., Mimoso, C.A., Adelman, K., and Neugebauer, K.M. (2021). Co-transcriptional splicing regulates 3′ end cleavage during mammalian erythropoiesis. Molecular Cell *81*, 998-1012.e7. 10.1016/j.molcel.2020.12.018.

2. Zhang, S., Aibara, S., Vos, S.M., Agafonov, D.E., Lührmann, R., and Cramer, P. (2021). Structure of a transcribing RNA polymerase II–U1 snRNP complex. Science *371*, 305–309. 10.1126/science.abf1870.

3. Herzel, L., Straube, K., and Neugebauer, K.M. (2018). Long-read sequencing of nascent RNA reveals coupling among RNA processing events. Genome Res. *28*, 1008–1019. 10.1101/gr.232025.117.

4. Herzel, L., Ottoz, D.S.M., Alpert, T., and Neugebauer, K.M. (2017). Splicing and transcription touch base: co-transcriptional spliceosome assembly and function. Nat Rev Mol Cell Biol *18*, 637–650. 10.1038/nrm.2017.63.

5. Caizzi, L., Monteiro-Martins, S., Schwalb, B., Lysakovskaia, K., Schmitzova, J., Sawicka, A., Chen, Y., Lidschreiber, M., and Cramer, P. (2021). Efficient RNA polymerase II pause release requires U2 snRNP function. Mol Cell *81*, 1920-1934.e9. 10.1016/j.molcel.2021.02.016.

6. Hluchý, M., Gajdušková, P., Ruiz de los Mozos, I., Rájecký, M., Kluge, M., Berger, B.-T., Slabá, Z., Potěšil, D., Weiß, E., Ule, J., et al. (2022). CDK11 regulates pre-mRNA splicing by phosphorylation of SF3B1. Nature *609*, 829–834. 10.1038/s41586-022-05204-z.

7. Fong, N., Kim, H., Zhou, Y., Ji, X., Qiu, J., Saldi, T., Diener, K., Jones, K., Fu, X.-D., and Bentley, D.L. (2014). Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate. Genes Dev. *28*, 2663–2676. 10.1101/gad.252106.114.

8. Inoue, D., Polaski, J.T., Taylor, J., Castel, P., Chen, S., Kobayashi, S., Hogg, S.J., Hayashi, Y., Pineda, J.M.B., El Marabti, E., et al. (2021). Minor intron retention drives clonal hematopoietic disorders and diverse cancer predisposition. Nat Genet *53*, 707–718. 10.1038/s41588-021-00828-9.

9. Yoshida, K., Sanada, M., Shiraishi, Y., Nowak, D., Nagata, Y., Yamamoto, R., Sato, Y., Sato-Otsubo, A., Kon, A., Nagasaki, M., et al. (2011). Frequent pathway mutations of splicing machinery in myelodysplasia. Nature *478*, 64–69. 10.1038/nature10496.

10. Hahn, C.N., and Scott, H.S. (2012). Spliceosome mutations in hematopoietic malignancies. Nat Genet *44*, 9–10. 10.1038/ng.1045.

11. Dvinge, H., Kim, E., Abdel-Wahab, O., and Bradley, R.K. (2016). RNA splicing factors as oncoproteins and tumour suppressors. Nat Rev Cancer *16*, 413–430. 10.1038/nrc.2016.51.

12. Sun, C. (2020). The SF3b complex: splicing and beyond. Cell. Mol. Life Sci. *77*, 3583–3595. 10.1007/s00018-020-03493-z.

13. Zhou, Z., Gong, Q., Wang, Y., Li, M., Wang, L., Ding, H., and Li, P. (2020). The biological function and clinical significance of SF3B1 mutations in cancer. Biomark Res *8*, 38. 10.1186/s40364-020-00220-5.

14. DeBoever, C., Ghia, E.M., Shepard, P.J., Rassenti, L., Barrett, C.L., Jepsen, K., Jamieson, C.H.M., Carson, D., Kipps, T.J., and Frazer, K.A. (2015). Transcriptome Sequencing Reveals Potential Mechanism of Cryptic 3’ Splice Site Selection in SF3B1-mutated Cancers. PLoS Comput Biol *11*, e1004105. 10.1371/journal.pcbi.1004105.

15. Zhang, J., Ali, A.M., Lieu, Y.K., Liu, Z., Gao, J., Rabadan, R., Raza, A., Mukherjee, S., and Manley, J.L. (2019). Disease-Causing Mutations in SF3B1 Alter Splicing by Disrupting Interaction with SUGP1. Molecular Cell *76*, 82-95.e7. 10.1016/j.molcel.2019.07.017.

16. Zhao, B., Li, Z., Qian, R., Liu, G., Fan, M., Liang, Z., Hu, X., and Wan, Y. (2022). Cancer-associated mutations in *SF3B1* disrupt the interaction between *SF3B1* and *DDX42*. The Journal of Biochemistry *172*, 117–126. 10.1093/jb/mvac049.

17. Clough, C.A., Pangallo, J., Sarchi, M., Ilagan, J.O., North, K., Bergantinos, R., Stolla, M.C., Naru, J., Nugent, P., Kim, E., et al. (2022). Coordinated missplicing of TMEM14C and ABCB7 causes ring sideroblast formation in SF3B1-mutant myelodysplastic syndrome. Blood *139*, 2038–2049. 10.1182/blood.2021012652.

18. Lieu, Y.K., Liu, Z., Ali, A.M., Wei, X., Penson, A., Zhang, J., An, X., Rabadan, R., Raza, A., Manley, J.L., et al. (2022). SF3B1 mutant-induced missplicing of MAP3K7 causes anemia in myelodysplastic syndromes. Proc. Natl. Acad. Sci. U.S.A. *119*, e2111703119. 10.1073/pnas.2111703119.

19. Inoue, D., Chew, G.-L., Liu, B., Michel, B.C., Pangallo, J., D'Avino, A.R., Hitchman, T., North, K., Lee, S.C.-W., Bitner, L., et al. (2019). Spliceosomal disruption of the non-canonical BAF complex in cancer. Nature *574*, 432–436. 10.1038/s41586-019-1646-9.

20. Kim, E., Ilagan, J.O., Liang, Y., Daubner, G.M., Lee, S.C.-W., Ramakrishnan, A., Li, Y., Chung, Y.R., Micol, J.-B., Murphy, M.E., et al. (2015). SRSF2 Mutations Contribute to Myelodysplasia by Mutant-Specific Effects on Exon Recognition. Cancer Cell *27*, 617–630. 10.1016/j.ccell.2015.04.006.

21. Boddu, P.C., Gupta, A.K., Kim, J.-S., Neugebauer, K.M., Waldman, T., and Pillai, M.M. (2021). Generation of scalable cancer models by combining AAV-intron-trap, CRISPR/Cas9, and inducible Cre-recombinase. Commun Biol *4*, 1184. 10.1038/s42003-021-02690-1.

22. Core, L.J., Waterfall, J.J., and Lis, J.T. (2008). Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. Science *322*, 1845–1848. 10.1126/science.1162228.

23. Reppas, N.B., Wade, J.T., Church, G.M., and Struhl, K. (2006). The Transition between Transcriptional Initiation and Elongation in E. coli Is Highly Variable and Often Rate Limiting. Molecular Cell *24*, 747–757. 10.1016/j.molcel.2006.10.030.

24. Barbieri, E., Hill, C., Quesnel-Vallières, M., Zucco, A.J., Barash, Y., and Gardini, A. (2020). Rapid and Scalable Profiling of Nascent RNA with fastGRO. Cell Reports *33*, 108373. 10.1016/j.celrep.2020.108373.

25. Ehrensberger, A.H., Kelly, G.P., and Svejstrup, J.Q. (2013). Mechanistic Interpretation of Promoter-Proximal Peaks and RNAPII Density Maps. Cell *154*, 713–715. 10.1016/j.cell.2013.07.032.

26. Booth, G.T., Wang, I.X., Cheung, V.G., and Lis, J.T. (2016). Divergence of a conserved elongation factor and transcription regulation in budding and fission yeast. Genome Res. *26*, 799–811. 10.1101/gr.204578.116.

27. Schofield, J.A., Duffy, E.E., Kiefer, L., Sullivan, M.C., and Simon, M.D. (2018). TimeLapse-seq: adding a temporal dimension to RNA sequencing through nucleoside recoding. Nat Methods *15*, 221–225. 10.1038/nmeth.4582.

28. Reimer, K.A., and Neugebauer, K.M. (2020). Preparation of Mammalian Nascent RNA for Long Read Sequencing. Current Protocols in Molecular Biology *133*. 10.1002/cpmb.128.

29. Chen, K., Hu, Z., Xia, Z., Zhao, D., Li, W., and Tyler, J.K. (2016). The Overlooked Fact: Fundamental Need for Spike-In Control for Virtually All Genome-Wide Analyses. Mol Cell Biol *36*, 662–667. 10.1128/MCB.00970-14.

30. Hershberger, C.E., Moyer, D.C., Adema, V., Kerr, C.M., Walter, W., Hutter, S., Meggendorfer, M., Baer, C., Kern, W., Nadarajah, N., et al. (2021). Complex landscape of alternative splicing in myeloid neoplasms. Leukemia *35*, 1108–1120. 10.1038/s41375-020-1002-y.

31. Middleton, R., Gao, D., Thomas, A., Singh, B., Au, A., Wong, J.J.-L., Bomane, A., Cosson, B., Eyras, E., Rasko, J.E.J., et al. (2017). IRFinder: assessing the impact of intron retention on mammalian gene expression. Genome Biol *18*, 51. 10.1186/s13059-017-1184-4.

32. Yoshimi, A., Lin, K.-T., Wiseman, D.H., Rahman, M.A., Pastore, A., Wang, B., Lee, S.C.-W., Micol, J.-B., Zhang, X.J., de Botton, S., et al. (2019). Coordinated alterations in RNA splicing and epigenetic regulation drive leukaemogenesis. Nature *574*, 273–277. 10.1038/s41586-019-1618-0.

33. Im, H., Rao, V., Sridhar, K., Bentley, J., Mishra, T., Chen, R., Hall, J., Graber, A., Zhang, Y., Li, X., et al. (2018). Distinct transcriptomic and exomic abnormalities within myelodysplastic syndrome marrow cells. Leukemia & Lymphoma *59*, 2952–2962. 10.1080/10428194.2018.1452210.

34. Shiozawa, Y., Malcovati, L., Gallì, A., Sato-Otsubo, A., Kataoka, K., Sato, Y., Watatani, Y., Suzuki, H., Yoshizato, T., Yoshida, K., et al. (2018). Aberrant splicing and defective mRNA production induced by somatic spliceosome mutations in myelodysplasia. Nat Commun *9*, 3649. 10.1038/s41467-018-06063-x.

35. Pellagatti, A., Armstrong, R.N., Steeples, V., Sharma, E., Repapi, E., Singh, S., Sanchi, A., Radujkovic, A., Horn, P., Dolatshad, H., et al. (2018). Impact of spliceosome mutations on RNA splicing in myelodysplasia: dysregulated genes/pathways and clinical associations. Blood *132*, 1225–1240. 10.1182/blood-2018-04-843771.

36. Zhang, Z., Rigo, N., Dybkov, O., Fourmann, J.-B., Will, C.L., Kumar, V., Urlaub, H., Stark, H., and Lührmann, R. (2021). Structural insights into how Prp5 proofreads the pre-mRNA branch site. Nature *596*, 296–300. 10.1038/s41586-021-03789-5.

37. Zhang, Z., Will, C.L., Bertram, K., Dybkov, O., Hartmuth, K., Agafonov, D.E., Hofele, R., Urlaub, H., Kastner, B., Lührmann, R., et al. (2020). Molecular architecture of the human 17S U2 snRNP. Nature *583*, 310–313. 10.1038/s41586-020-2344-3.

38. Tang, Q., Rodriguez-Santiago, S., Wang, J., Pu, J., Yuste, A., Gupta, V., Moldón, A., Xu, Y.-Z., and Query, C.C. (2016). SF3B1/Hsh155 HEAT motif mutations affect interaction with the spliceosomal ATPase Prp5, resulting in altered branch site selectivity in pre-mRNA splicing. Genes Dev. *30*, 2710–2723. 10.1101/gad.291872.116.

39. Kfir, N., Lev-Maor, G., Glaich, O., Alajem, A., Datta, A., Sze, S.K., Meshorer, E., and Ast, G. (2015). SF3B1 Association with Chromatin Determines Splicing Outcomes. Cell Reports *11*, 618–629. 10.1016/j.celrep.2015.03.048.

40. Carrocci, T.J., and Neugebauer, K.M. (2019). Pre-mRNA Splicing in the Nuclear Landscape. Cold Spring Harb Symp Quant Biol *84*, 11–20. 10.1101/sqb.2019.84.040402.

41. Castillo-Guzman, D., Hartono, S.R., Sanz, L.A., and Chédin, F. (2020). SF3B1-targeted Splicing Inhibition Triggers Global Alterations in Transcriptional Dynamics and R-Loop Metabolism. 2020.06.08.130583. 10.1101/2020.06.08.130583.

42. Lin, C., Smith, E.R., Takahashi, H., Lai, K.C., Martin-Brown, S., Florens, L., Washburn, M.P., Conaway, J.W., Conaway, R.C., and Shilatifard, A. (2010). AFF4, a Component of the ELL/P-TEFb Elongation Complex and a Shared Subunit of MLL Chimeras, Can Link Transcription Elongation to Leukemia. Molecular Cell *37*, 429–437. 10.1016/j.molcel.2010.01.026.

43. Luo, Z., Lin, C., and Shilatifard, A. (2012). The super elongation complex (SEC) family in transcriptional control. Nat Rev Mol Cell Biol *13*, 543–547. 10.1038/nrm3417.

44. Castillo-Guzman, D., Hartono, S.R., Sanz, L.A., and Chédin, F. (2020). SF3B1-targeted Splicing Inhibition Triggers Global Alterations in Transcriptional Dynamics and R-Loop Metabolism (Molecular Biology) 10.1101/2020.06.08.130583.

45. Rosa-Mercado, N.A., Zimmer, J.T., Apostolidi, M., Rinehart, J., Simon, M.D., and Steitz, J.A. (2021). Hyperosmotic stress alters the RNA polymerase II interactome and induces readthrough transcription despite widespread transcriptional repression. Molecular Cell *81*, 502-513.e4. 10.1016/j.molcel.2020.12.002.

46. Chen, L., Chen, J.-Y., Huang, Y.-J., Gu, Y., Qiu, J., Qian, H., Shao, C., Zhang, X., Hu, J., Li, H., et al. (2018). The Augmented R-Loop Is a Unifying Mechanism for Myelodysplastic Syndromes Induced by High-Risk Splicing Factor Mutations. Molecular Cell *69*, 412-425.e6. 10.1016/j.molcel.2017.12.029.

47. Bapat, A., Keita, N., Martelly, W., Kang, P., Seet, C., Jacobsen, J.R., Stoilov, P., Hu, C., Crooks, G.M., and Sharma, S. (2018). Myeloid Disease Mutations of Splicing Factor SRSF2 Cause G2-M Arrest and Skewed Differentiation of Human Hematopoietic Stem and Progenitor Cells. Stem Cells *36*, 1663–1675. 10.1002/stem.2885.

48. Kim, S., Park, C., Jun, Y., Lee, S., Jung, Y., and Kim, J. (2018). Integrative Profiling of Alternative Splicing Induced by U2AF1 S34F Mutation in Lung Adenocarcinoma Reveals a Mechanistic Link to Mitotic Stress. Mol Cells *41*, 733–741. 10.14348/molcells.2018.0176.

49. Singh, S., Ahmed, D., Dolatshad, H., Tatwavedi, D., Schulze, U., Sanchi, A., Ryley, S., Dhir, A., Carpenter, L., Watt, S.M., et al. (2020). SF3B1 mutations induce R-loop accumulation and DNA damage in MDS and leukemia cells with therapeutic implications. Leukemia *34*, 2525–2530. 10.1038/s41375-020-0753-9.

50. Smolka, J.A., Sanz, L.A., Hartono, S.R., and Chédin, F. (2021). Recognition of RNA by the S9.6 antibody creates pervasive artifacts when imaging RNA:DNA hybrids. Journal of Cell Biology *220*, e202004079. 10.1083/jcb.202004079.

51. Bayona-Feliu, A., Barroso, S., Muñoz, S., and Aguilera, A. (2021). The SWI/SNF chromatin remodeling complex helps resolve R-loop-mediated transcription–replication conflicts. Nat Genet *53*, 1050–1063. 10.1038/s41588-021-00867-2.

52. Hegazy, M., Cohen-Barak, E., Koetsier, J.L., Najor, N.A., Arvanitis, C., Sprecher, E., Green, K.J., and Godsel, L.M. (2020). Proximity Ligation Assay for Detecting Protein-Protein Interactions and Protein Modifications in Cells and Tissues in Situ. Current Protocols in Cell Biology *89*. 10.1002/cpcb.115.

53. Karlsson, A., Deb-Basu, D., Cherry, A., Turner, S., Ford, J., and Felsher, D.W. (2003). Defective double-strand DNA break repair and chromosomal translocations by *MYC* overexpression. Proc. Natl. Acad. Sci. U.S.A. *100*, 9974–9979. 10.1073/pnas.1732638100.

54. Zeman, M.K., and Cimprich, K.A. (2014). Causes and consequences of replication stress. Nat Cell Biol *16*, 2–9. 10.1038/ncb2897.

55. Pérez-Calero, C., Bayona-Feliu, A., Xue, X., Barroso, S.I., Muñoz, S., González-Basallote, V.M., Sung, P., and Aguilera, A. (2020). UAP56/DDX39B is a major cotranscriptional RNA–DNA helicase that unwinds harmful R loops genome-wide. Genes Dev. *34*, 898–912. 10.1101/gad.336024.119.

56. Porrua, O., and Libri, D. (2015). Transcription termination and the control of the transcriptome: why, where and how to stop. Nat Rev Mol Cell Biol *16*, 190–202. 10.1038/nrm3943.

57. Okamoto, Y., Abe, M., Itaya, A., Tomida, J., Ishiai, M., Takaori-Kondo, A., Taoka, M., Isobe, T., and Takata, M. (2019). FANCD 2 protects genome stability by recruiting RNA processing enzymes to resolve R-loops during mild replication stress. FEBS J *286*, 139–150. 10.1111/febs.14700.

58. García-Muse, T., and Aguilera, A. (2019). R Loops: From Physiological to Pathological Roles. Cell *179*, 604–618. 10.1016/j.cell.2019.08.055.

59. Rinaldi, C., Pizzul, P., Longhese, M.P., and Bonetti, D. (2021). Sensing R-Loop-Associated DNA Damage to Safeguard Genome Stability. Front. Cell Dev. Biol. *8*, 618157. 10.3389/fcell.2020.618157.

60. Kemiha, S., Poli, J., Lin, Y.-L., Lengronne, A., and Pasero, P. (2021). Toxic R-loops: Cause or consequence of replication stress? DNA Repair *107*, 103199. 10.1016/j.dnarep.2021.103199.

61. Crossley, M.P., Bocek, M., and Cimprich, K.A. (2019). R-Loops as Cellular Regulators and Genomic Threats. Molecular Cell *73*, 398–411. 10.1016/j.molcel.2019.01.024.

62. Bensaude, O. (2011). Inhibiting eukaryotic transcription. Which compound to choose? How to evaluate its activity?: Which compound to choose? How to evaluate its activity? Transcription *2*, 103–108. 10.4161/trns.2.3.16172.

63. Muralimanoharan, S., Shamby, R., Stansbury, N., Schenken, R., de la Pena Avalos, B., Javanmardi, S., Dray, E., Sung, P., and Boyer, T.G. (2022). Aberrant R-loop-induced replication stress in MED12-mutant uterine fibroids. Sci Rep *12*, 6169. 10.1038/s41598-022-10188-x.

64. Flomerfelt, F.A., and Gress, R.E. (2016). Analysis of Cell Proliferation and Homeostasis Using EdU Labeling. In T-Cell Development Methods in Molecular Biology., R. Bosselut and M. S. Vacchio, eds. (Springer New York), pp. 211–220. 10.1007/978-1-4939-2809-5_18.

65. Zatreanu, D., Han, Z., Mitter, R., Tumini, E., Williams, H., Gregersen, L., Dirac-Svejstrup, A.B., Roma, S., Stewart, A., Aguilera, A., et al. (2019). Elongation Factor TFIIS Prevents Transcription Stress and R-Loop Accumulation to Maintain Genome Stability. Molecular Cell *76*, 57-69.e9. 10.1016/j.molcel.2019.07.037.

66. Castillo-Guzman, D., and Chédin, F. (2021). Defining R-loop classes and their contributions to genome instability. DNA Repair *106*, 103182. 10.1016/j.dnarep.2021.103182.

67. Hamperl, S., Bocek, M.J., Saldivar, J.C., Swigut, T., and Cimprich, K.A. (2017). Transcription-Replication Conflict Orientation Modulates R-Loop Levels and Activates Distinct DNA Damage Responses. Cell *170*, 774-786.e19. 10.1016/j.cell.2017.07.043.

68. Sanz, L.A., and Chédin, F. (2019). High-resolution, strand-specific R-loop mapping via S9.6-based DNA–RNA immunoprecipitation and high-throughput sequencing. Nat Protoc *14*, 1734–1755. 10.1038/s41596-019-0159-1.

69. Promonet, A., Padioleau, I., Liu, Y., Sanz, L., Biernacka, A., Schmitz, A.-L., Skrzypczak, M., Sarrazin, A., Mettling, C., Rowicka, M., et al. (2020). Topoisomerase 1 prevents replication stress at R-loop-enriched transcription termination sites. Nat Commun *11*, 3940. 10.1038/s41467-020-17858-2.

70. Pongor, L.S., Gross, J.M., Vera Alvarez, R., Murai, J., Jang, S.-M., Zhang, H., Redon, C., Fu, H., Huang, S.-Y., Thakur, B., et al. (2020). BAMscale: quantification of next-generation sequencing peaks and generation of scaled coverage tracks. Epigenetics & Chromatin *13*, 21. 10.1186/s13072-020-00343-x.

71. Al-Hadid, Q., and Yang, Y. (2016). R-loop: an emerging regulator of chromatin dynamics: Table 1. Acta Biochim Biophys Sin *48*, 623–631. 10.1093/abbs/gmw052.

72. Li, B., Carey, M., and Workman, J.L. (2007). The Role of Chromatin during Transcription. Cell *128*, 707–719. 10.1016/j.cell.2007.01.015.

73. Kuendgen, A., Müller-Thomas, C., Lauseker, M., Haferlach, T., Urbaniak, P., Schroeder, T., Brings, C., Wulfert, M., Meggendorfer, M., Hildebrandt, B., et al. (2018). Efficacy of azacitidine is independent of molecular and clinical characteristics - an analysis of 128 patients with myelodysplastic syndromes or acute myeloid leukemia and a review of the literature. Oncotarget *9*, 27882–27894. 10.18632/oncotarget.25328.

74. Schroeder, M.A., and DeZern, A.E. (2015). Do somatic mutations in de novo MDS predict for response to treatment? Hematology *2015*, 317–328. 10.1182/asheducation-2015.1.317.

75. Skene, P.J., and Henikoff, S. (2017). An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. eLife *6*, e21856. 10.7554/eLife.21856.

76. Bieberstein, N.I., Carrillo Oesterreich, F., Straube, K., and Neugebauer, K.M. (2012). First Exon Length Controls Active Chromatin Signatures and Transcription. Cell Reports *2*, 62–68. 10.1016/j.celrep.2012.05.019.

77. Wang, Z., Chivu, A.G., Choate, L.A., Rice, E.J., Miller, D.C., Chu, T., Chou, S.-P., Kingsley, N.B., Petersen, J.L., Finno, C.J., et al. (2020). Interdependence between histone marks and steps in Pol II transcription (Genomics) 10.1101/2020.04.08.032730.

78. Le Martelot, G., Canella, D., Symul, L., Migliavacca, E., Gilardi, F., Liechti, R., Martin, O., Harshman, K., Delorenzi, M., Desvergne, B., et al. (2012). Genome-Wide RNA Polymerase II Profiles and RNA Accumulation Reveal Kinetics of Transcription and Associated Epigenetic Changes During Diurnal Cycles. PLoS Biol *10*, e1001442. 10.1371/journal.pbio.1001442.

79. Mas, G., Blanco, E., Ballaré, C., Sansó, M., Spill, Y.G., Hu, D., Aoi, Y., Le Dily, F., Shilatifard, A., Marti-Renom, M.A., et al. (2018). Promoter bivalency favors an open chromatin architecture in embryonic stem cells. Nat Genet *50*, 1452–1462. 10.1038/s41588-018-0218-5.

80. Fazzio, T.G. (2016). Regulation of chromatin structure and cell fate by R-loops. Transcription *7*, 121–126. 10.1080/21541264.2016.1198298.

81. Gel, B., Díez-Villanueva, A., Serra, E., Buschbeck, M., Peinado, M.A., and Malinverni, R. (2015). regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. Bioinformatics, btv562. 10.1093/bioinformatics/btv562.

82. Schep, A.N., Buenrostro, J.D., Denny, S.K., Schwartz, K., Sherlock, G., and Greenleaf, W.J. (2015). Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. Genome Res. *25*, 1757–1770. 10.1101/gr.192294.115.

83. Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thåström, A., Field, Y., Moore, I.K., Wang, J.-P.Z., and Widom, J. (2006). A genomic code for nucleosome positioning. Nature *442*, 772–778. 10.1038/nature04979.

84. Parikh, R.Y., and Kim, H.D. (2013). The Effect of an Intervening Promoter Nucleosome on Gene Expression. PLoS ONE *8*, e63072. 10.1371/journal.pone.0063072.

85. Hu, S., Peng, L., Xu, C., Wang, Z., Song, A., and Chen, F.X. (2021). SPT5 stabilizes RNA polymerase II, orchestrates transcription cycles, and maintains the enhancer landscape. Molecular Cell *81*, 4425-4439.e6. 10.1016/j.molcel.2021.08.029.

86. Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. Nat Methods *9*, 215–216. 10.1038/nmeth.1906.

87. Schneider, J., Wood, A., Lee, J.-S., Schuster, R., Dueker, J., Maguire, C., Swanson, S.K., Florens, L., Washburn, M.P., and Shilatifard, A. (2005). Molecular Regulation of Histone H3 Trimethylation by COMPASS and the Regulation of Gene Expression. Molecular Cell *19*, 849–856. 10.1016/j.molcel.2005.07.024.

88. Wysocka, J., Myers, M.P., Laherty, C.D., Eisenman, R.N., and Herr, W. (2003). Human Sin3 deacetylase and trithorax-related Set1/Ash2 histone H3-K4 methyltransferase are tethered together selectively by the cell-proliferation factor HCF-1. Genes Dev. *17*, 896–911. 10.1101/gad.252103.

89. Beurton, F., Stempor, P., Caron, M., Appert, A., Dong, Y., Chen, R.A., Cluet, D., Couté, Y., Herbette, M., Huang, N., et al. (2019). Physical and functional interaction between SET1/COMPASS complex component CFP-1 and a Sin3S HDAC complex in C. elegans. Nucleic Acids Research *47*, 11164–11180. 10.1093/nar/gkz880.

90. Yang, Y., Huang, W., Qiu, R., Liu, R., Zeng, Y., Gao, J., Zheng, Y., Hou, Y., Wang, S., Yu, W., et al. (2018). LSD1 coordinates with the SIN3A/HDAC complex and maintains sensitivity to chemotherapy in breast cancer. Journal of Molecular Cell Biology *10*, 285–301. 10.1093/jmcb/mjy021.

91. Banks, C.A.S., Zhang, Y., Miah, S., Hao, Y., Adams, M.K., Wen, Z., Thornton, J.L., Florens, L., and Washburn, M.P. (2020). Integrative Modeling of a Sin3/HDAC Complex Sub-structure. Cell Reports *31*, 107516. 10.1016/j.celrep.2020.03.080.

92. Liu, M., and Pile, L.A. (2017). The Transcriptional Corepressor SIN3 Directly Regulates Genes Involved in Methionine Catabolism and Affects Histone Methylation, Linking Epigenetics and Metabolism. Journal of Biological Chemistry *292*, 1970–1976. 10.1074/jbc.M116.749754.

93. Iwase, S., Januma, A., Miyamoto, K., Shono, N., Honda, A., Yanagisawa, J., and Baba, T. (2004). Characterization of BHC80 in BRAF–HDAC complex, involved in neuron-specific gene repression. Biochemical and Biophysical Research Communications *322*, 601–608. 10.1016/j.bbrc.2004.07.163.

94. Iwase, S., Shono, N., Honda, A., Nakanishi, T., Kashiwabara, S., Takahashi, S., and Baba, T. (2006). A component of BRAF-HDAC complex, BHC80, is required for neonatal survival in mice. FEBS Letters *580*, 3129–3135. 10.1016/j.febslet.2006.04.065.

95. Zhang, T., Cooper, S., and Brockdorff, N. (2015). The interplay of histone modifications – writers that read. EMBO Rep *16*, 1467–1481. 10.15252/embr.201540945.

96. Carugo, A., Genovese, G., Seth, S., Nezi, L., Rose, J.L., Bossi, D., Cicalese, A., Shah, P.K., Viale, A., Pettazzoni, P.F., et al. (2016). In Vivo Functional Platform Targeting Patient-Derived Xenografts Identifies WDR5-Myc Association as a Critical Determinant of Pancreatic Cancer. Cell Reports *16*, 133–147. 10.1016/j.celrep.2016.05.063.

97. Neilsen, B.K., Chakraborty, B., McCall, J.L., Frodyma, D.E., Sleightholm, R.L., Fisher, K.W., and Lewis, R.E. (2018). WDR5 supports colon cancer cells by promoting methylation of H3K4 and suppressing DNA damage. BMC Cancer *18*, 673. 10.1186/s12885-018-4580-6.

98. Zhou, Q., Chen, X., He, H., Peng, S., Zhang, Y., Zhang, J., Cheng, L., Liu, S., Huang, M., Xie, R., et al. (2021). WD repeat domain 5 promotes chemoresistance and Programmed Death-Ligand 1 expression in prostate cancer. Theranostics *11*, 4809–4824. 10.7150/thno.55814.

99. Grebien, F., Vedadi, M., Getlik, M., Giambruno, R., Grover, A., Avellino, R., Skucha, A., Vittori, S., Kuznetsova, E., Smil, D., et al. (2015). Pharmacological targeting of the Wdr5-MLL interaction in C/EBPα N-terminal leukemia. Nat Chem Biol *11*, 571–578. 10.1038/nchembio.1859.

100. Obeng, E.A., Chappell, R.J., Seiler, M., Chen, M.C., Campagna, D.R., Schmidt, P.J., Schneider, R.K., Lord, A.M., Wang, L., Gambe, R.G., et al. (2016). Physiologic Expression of Sf3b1 K700E Causes Impaired Erythropoiesis, Aberrant Splicing, and Sensitivity to Therapeutic Spliceosome Modulation. Cancer Cell *30*, 404–417. 10.1016/j.ccell.2016.08.006.

101. Stanley, R.F., and Abdel-Wahab, O. (2022). Dysregulation and therapeutic targeting of RNA splicing in cancer. Nat Cancer *3*, 536–546. 10.1038/s43018-022-00384-z.

102. Bradley, R.K., and Anczuków, O. (2023). RNA splicing dysregulation and the hallmarks of cancer. Nat Rev Cancer *23*, 135–155. 10.1038/s41568-022-00541-7.

103. Chédin, F. (2016). Nascent Connections: R-Loops and Chromatin Patterning. Trends in Genetics *32*, 828–838. 10.1016/j.tig.2016.10.002.

104. Venkatesh, S., and Workman, J.L. (2015). Histone exchange, chromatin structure and the regulation of transcription. Nat Rev Mol Cell Biol *16*, 178–189. 10.1038/nrm3941.

105. Papaemmanuil, E., Gerstung, M., Malcovati, L., Tauro, S., Gundem, G., Van Loo, P., Yoon, C.J., Ellis, P., Wedge, D.C., Pellagatti, A., et al. (2013). Clinical and biological implications of driver mutations in myelodysplastic syndromes. Blood *122*, 3616–3627. 10.1182/blood-2013-08-518886.

106. Haferlach, T., Nagata, Y., Grossmann, V., Okuno, Y., Bacher, U., Nagae, G., Schnittger, S., Sanada, M., Kon, A., Alpermann, T., et al. (2014). Landscape of genetic lesions in 944 patients with myelodysplastic syndromes. Leukemia *28*, 241–247. 10.1038/leu.2013.336.

107. Xu, J., Song, F., Lyu, H., Kobayashi, M., Zhang, B., Zhao, Z., Hou, Y., Wang, X., Luan, Y., Jia, B., et al. (2022). Subtype-specific 3D genome alteration in acute myeloid leukaemia. Nature *611*, 387–398. 10.1038/s41586-022-05365-x.

108. Shilatifard, A. (2012). The COMPASS Family of Histone H3K4 Methylases: Mechanisms of Regulation in Development and Disease Pathogenesis. Annu. Rev. Biochem. *81*, 65–95. 10.1146/annurev-biochem-051710-134100.

109. Chathoth, K.T., Barrass, J.D., Webb, S., and Beggs, J.D. (2014). A Splicing-Dependent Transcriptional Checkpoint Associated with Prespliceosome Formation. Molecular Cell *53*, 779–790. 10.1016/j.molcel.2014.01.017.

110. Loerch, S., Leach, J.R., Horner, S.W., Maji, D., Jenkins, J.L., Pulvino, M.J., and Kielkopf, C.L. (2019). The pre-mRNA splicing and transcription factor Tat-SF1 is a functional partner of the spliceosome SF3b1 subunit via a U2AF homology motif interface. Journal of Biological Chemistry *294*, 2892–5793. 10.1074/jbc.RA118.006764.

111. Talkish, J., Igel, H., Hunter, O., Horner, S.W., Jeffery, N.N., Leach, J.R., Jenkins, J.L., Kielkopf, C.L., and Ares, M. (2019). Cus2 enforces the first ATP-dependent step of splicing by binding to yeast SF3b1 through a UHM–ULM interaction. RNA *25*, 1020–1037. 10.1261/rna.070649.119.

112. Will, C.L., and Luhrmann, R. (2011). Spliceosome Structure and Function. Cold Spring Harbor Perspectives in Biology *3*, a003707–a003707. 10.1101/cshperspect.a003707.

113. Gillotin, S. (2018). Isolation of Chromatin-bound Proteins from Subcellular Fractions for Biochemical Analysis. BIO-PROTOCOL *8*. 10.21769/BioProtoc.3035.

114. Folco, E.G., Lei, H., Hsu, J.L., and Reed, R. (2012). Small-scale Nuclear Extracts for Functional Assays of Gene-expression Machineries. JoVE, 4140. 10.3791/4140.

115. Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., et al. (2012). Fiji: an open-source platform for biological-image analysis. Nat Methods *9*, 676–682. 10.1038/nmeth.2019.

116. Boutet-Robinet, E., Trouche, D., and Canitrot, Y. (2013). Neutral Comet Assay. BIO-PROTOCOL *3*. 10.21769/BioProtoc.915.

117. Lam, F.C., Kong, Y.W., Huang, Q., Vu Han, T.-L., Maffa, A.D., Kasper, E.M., and Yaffe, M.B. (2020). BRD4 prevents the accumulation of R-loops and protects against transcription–replication collision events and DNA damage. Nat Commun *11*, 4083. 10.1038/s41467-020-17503-y.

118. Carey, M.F., Peterson, C.L., and Smale, S.T. (2009). Chromatin Immunoprecipitation (ChIP). Cold Spring Harb Protoc *2009*, pdb.prot5279. 10.1101/pdb.prot5279.

119. Egan, B., Yuan, C.-C., Craske, M.L., Labhart, P., Guler, G.D., Arnott, D., Maile, T.M., Busby, J., Henry, C., Kelly, T.K., et al. (2016). An Alternative Approach to ChIP-Seq Normalization Enables Detection of Genome-Wide Changes in Histone H3 Lysine 27 Trimethylation upon EZH2 Inhibition. PLoS ONE *11*, e0166438. 10.1371/journal.pone.0166438.

120. Skene, P.J., Henikoff, J.G., and Henikoff, S. (2018). Targeted in situ genome-wide profiling with high efficiency for low cell numbers. Nat Protoc *13*, 1006–1019. 10.1038/nprot.2018.015.

121. Corces, M.R., Trevino, A.E., Hamilton, E.G., Greenside, P.G., Sinnott-Armstrong, N.A., Vesuna, S., Satpathy, A.T., Rubin, A.J., Montine, K.S., Wu, B., et al. (2017). An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. Nat Methods *14*, 959–962. 10.1038/nmeth.4396.

122. Baldwin, A., Morris, A.R., and Mukherjee, N. (2021). An Easy, Cost-Effective, and Scalable Method to Deplete Human Ribosomal RNA for RNA-seq. Current Protocols *1*. 10.1002/cpz1.176.

123. Wingett, S.W., and Andrews, S. (2018). FastQ Screen: A tool for multi-genome mapping and quality control. F1000Res *7*, 1338. 10.12688/f1000research.15931.2.

124. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet j. *17*, 10. 10.14806/ej.17.1.200.

125. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357–359. 10.1038/nmeth.1923.

126. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21. 10.1093/bioinformatics/bts635.

127. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. 10.1093/bioinformatics/btp352.

128. Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res *44*, W160–W165. 10.1093/nar/gkw257.

129. Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat Biotechnol *29*, 24–26. 10.1038/nbt.1754.

130. Hahne, F., and Ivanek, R. (2016). Visualizing Genomic Data Using Gviz and Bioconductor. In Statistical Genomics Methods in Molecular Biology., E. Mathé and S. Davis, eds. (Springer New York), pp. 335–351. 10.1007/978-1-4939-3578-9_16.

131. Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis.

132. Shen, S., Park, J.W., Lu, Z., Lin, L., Henry, M.D., Wu, Y.N., Zhou, Q., and Xing, Y. (2014). rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proc. Natl. Acad. Sci. U.S.A. *111*. 10.1073/pnas.1419161111.

133. Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930. 10.1093/bioinformatics/btt656.

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

135. Stovner, E.B., and Sætrom, P. (2019). epic2 efficiently finds diffuse domains in ChIP-seq data. Bioinformatics *35*, 4392–4393. 10.1093/bioinformatics/btz232.

136. Duttke, S.H., Chang, M.W., Heinz, S., and Benner, C. (2019). Identification and dynamic quantification of regulatory elements using total RNA. Genome Res. *29*, 1836–1846. 10.1101/gr.253492.119.

137. Wu, X., Kabalane, H., Kahli, M., Petryk, N., Laperrousaz, B., Jaszczyszyn, Y., Drillon, G., Nicolini, F.-E., Perot, G., Robert, A., et al. (2018). Developmental and cancer-associated plasticity of DNA replication preferentially targets GC-poor, lowly expressed and late-replicating regions. Nucleic Acids Research *46*, 10157–10172. 10.1093/nar/gky797.

138. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. 10.1093/bioinformatics/btq033.

139. Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. Bioinformatics *27*, 863–864. 10.1093/bioinformatics/btr026.

140. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100. 10.1093/bioinformatics/bty191.

141. Meers, M.P., Tenenbaum, D., and Henikoff, S. (2019). Peak calling by Sparse Enrichment Analysis for CUT&RUN chromatin profiling. Epigenetics & Chromatin *12*, 42. 10.1186/s13072-019-0287-4.

142. Zhang, Y., Park, C., Bennett, C., Thornton, M., and Kim, D. (2021). Rapid and accurate alignment of nucleotide conversion sequencing reads with HISAT-3N. Genome Res. *31*, 1290–1295. 10.1101/gr.275193.120.

143. Roth, S.J., Heinz, S., and Benner, C. (2020). ARTDeco: automatic readthrough transcription detection. BMC Bioinformatics *21*, 214. 10.1186/s12859-020-03551-0.

144. Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics *32*, 2847–2849. 10.1093/bioinformatics/btw313.

145. Moffat, J., Grueneberg, D.A., Yang, X., Kim, S.Y., Kloepfer, A.M., Hinkle, G., Piqani, B., Eisenhaure, T.M., Luo, B., Grenier, J.K., et al. (2006). A Lentiviral RNAi Library for Human and Mouse Genes Applied to an Arrayed Viral High-Content Screen. Cell *124*, 1283–1298. 10.1016/j.cell.2006.01.040.

146. Li, W., Xu, H., Xiao, T., Cong, L., Love, M.I., Zhang, F., Irizarry, R.A., Liu, J.S., Brown, M., and Liu, X.S. (2014). MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. Genome Biol *15*, 554. 10.1186/s13059-014-0554-4.

147. Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., et al. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. The Innovation *2*, 100141. 10.1016/j.xinn.2021.100141.

148. Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P., et al. (2019). STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Research *47*, D607–D613. 10.1093/nar/gky1131.

FIGURE 1

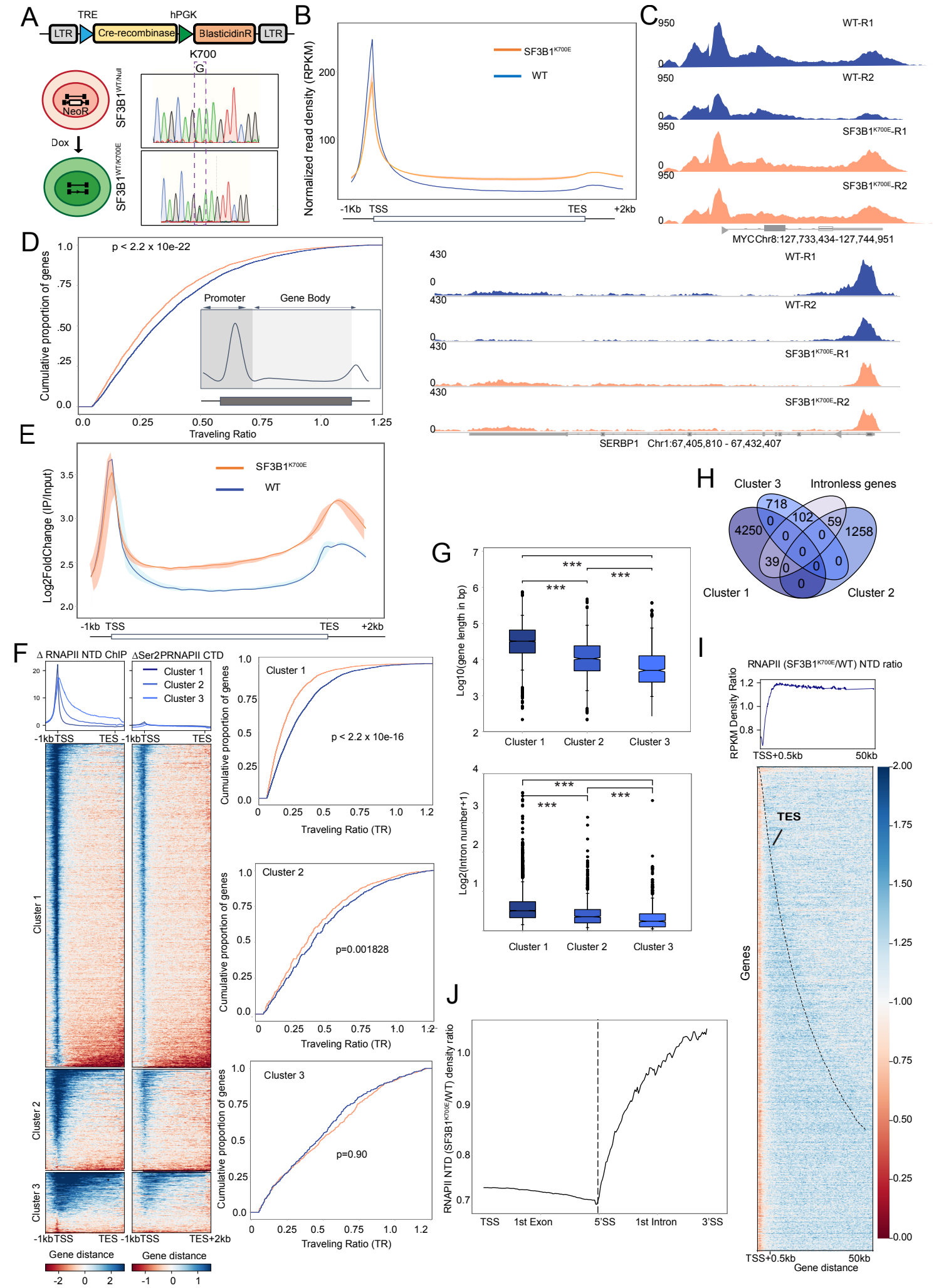

A
TRE
hPGK
LTR
Cre-recombinase
BlasticidinR
LTR
K700
G
NeoR
Dox
SF3B1WT/Null
SF3B1WT/K700E
B
Normalized read density (RPKM)
SF3B1K700E
WT
-1Kb TSS
TES
+2kb
C
WT-R1
WT-R2
SF3B1K700E-R1
SF3B1K700E-R2
MYC Chr8:127,733,434-127,744,951
SERBP1 Chr1:67,405,810 - 67,432,407
D
p < 2.2 x 10e-22
Cumulative proportion of genes
Promoter
Gene Body
Traveling Ratio
E
Log2FoldChange (IP/Input)
SF3B1K700E
WT
-1kb TSS
TES
+2kb
F
Δ RNAPII NTD ChIP
ΔSer2PRNAPII CTD
Cluster 1
Cluster 2
Cluster 3
-1kbTSS
TES
+2kb
Gene distance
Cumulative proportion of genes
p < 2.2 x 10e-16
p=0.001828
p=0.90
Traveling Ratio (TR)
G
Log10(gene length in bp)
Log2(Intron number+1)
***
H
Cluster 3
Intronless genes
Cluster 1
Cluster 2
718
102
59
4250
1258
39
I
RNAPII (SF3B1K700E/WT) NTD ratio
RPKM Density Ratio
TSS+0.5kb
50kb
TES
Genes
Gene distance
J
RNAPII NTD (SF3B1K700E/WT) density ratio
TSS
1st Exon
5'SS
1st Intron
3'SS

FIGURE 2

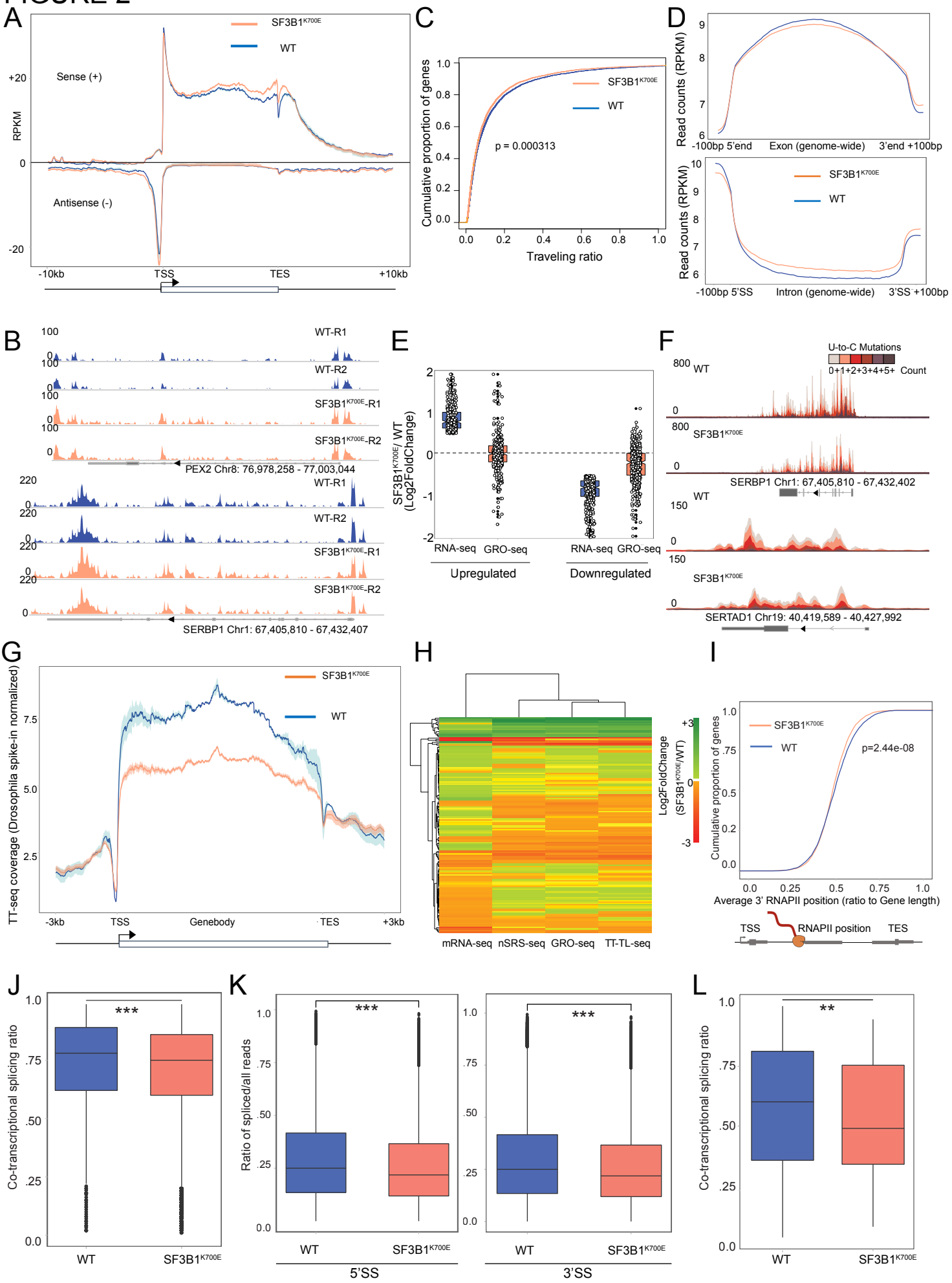

A
SF3B1K700E
WT
Sense (+)
Antisense (-)
RPKM
+20
0
-20
-10kb
TSS
TES
+10kb
C
Cumulative proportion of genes
Traveling ratio
p = 0.000313
D
Read counts (RPKM)
-100bp 5'end
Exon (genome-wide)
3'end +100bp
-100bp 5'SS
Intron (genome-wide)
3'SS +100bp
B
WT-R1
WT-R2
SF3B1K700E-R1
SF3B1K700E-R2
PEX2 Chr8: 76,978,258 - 77,003,044
SERBP1 Chr1: 67,405,810 - 67,432,407
E
SF3B1K700E/ WT (Log2FoldChange)
RNA-seq
GRO-seq
Upregulated
Downregulated
F
U-to-C Mutations
0+1+2+3+4+5+ Count
SERBP1 Chr1: 67,405,810 - 67,432,402
SERTAD1 Chr19: 40,419,589 - 40,427,992
G
TT-seq coverage (Drosophila spike-in normalized)
-3kb
TSS
Genebody
TES
+3kb
H
Log2FoldChange (SF3B1K700E/WT)
mRNA-seq
nSRS-seq
GRO-seq
TT-TL-seq
I
p=2.44e-08
Cumulative proportion of genes
Average 3' RNAPII position (ratio to Gene length)
RNAPII position
J
Co-transcriptional splicing ratio
***
K
Ratio of spliced/all reads
5'SS
3'SS
L
**

FIGURE 3

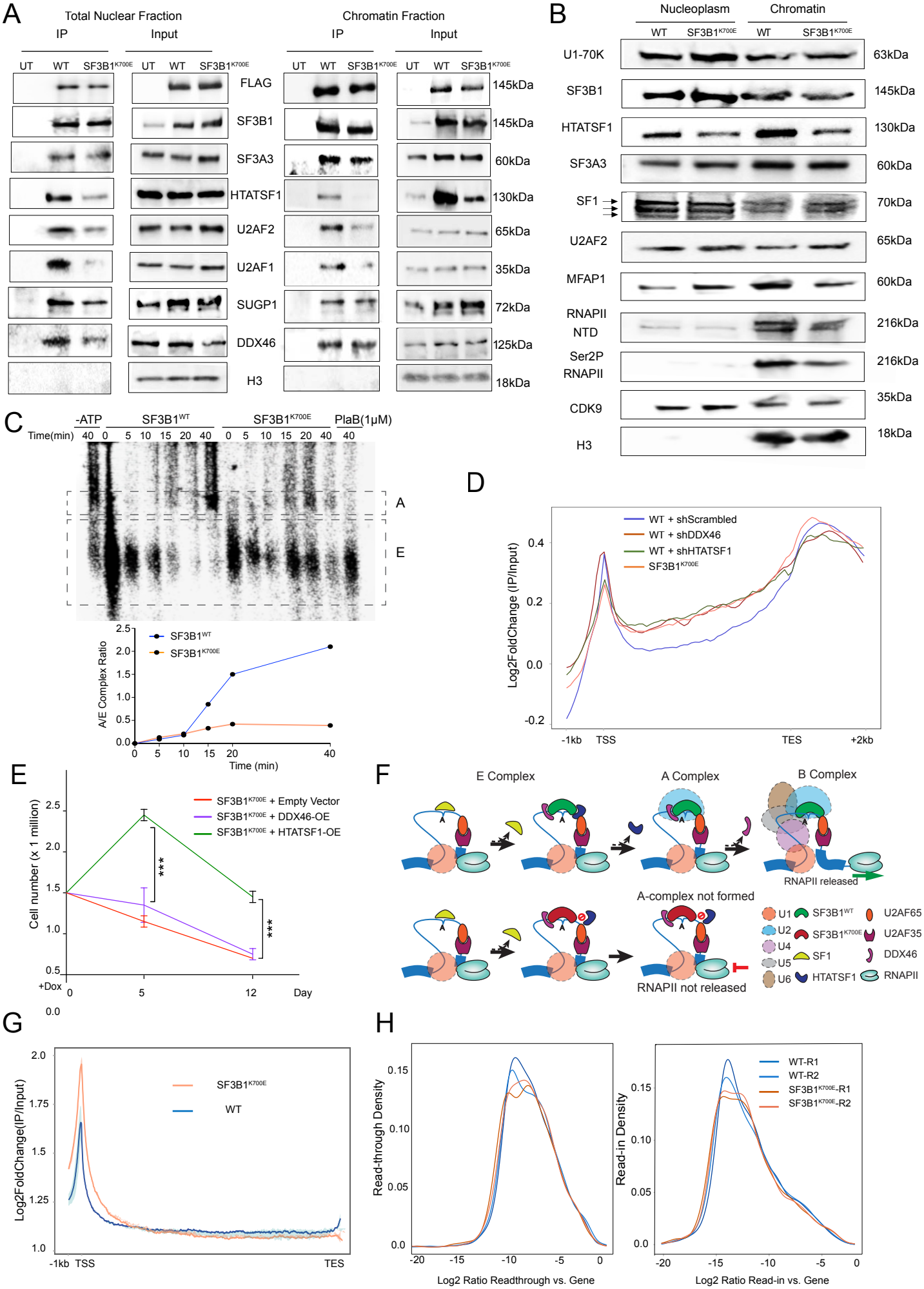

A
Total Nuclear Fraction
Chromatin Fraction
IP
Input
UT
WT
SF3B1K700E
FLAG
SF3B1
SF3A3
HTATSF1
U2AF2
U2AF1
SUGP1
DDX46
H3
145kDa
60kDa
130kDa
65kDa
35kDa
72kDa
125kDa
18kDa
B
Nucleoplasm
Chromatin
U1-70K
SF1
MFAP1
RNAPII NTD
Ser2P RNAPII
CDK9
63kDa
70kDa
216kDa
C
-ATP
PlaB(1μM)
Time(min)
A
E
A/E Complex Ratio
Time (min)
D
WT + shScrambled
WT + shDDX46
WT + shHTATSF1
Log2FoldChange (IP/Input)
-1kb
TSS
TES
+2kb
E
SF3B1K700E + Empty Vector
SF3B1K700E + DDX46-OE
SF3B1K700E + HTATSF1-OE
Cell number (x 1 million)
+Dox
Day
F
E Complex
A Complex
B Complex
RNAPII released
A-complex not formed
RNAPII not released
U1
U2
U4
U5
U6
SF1
HTATSF1
U2AF65
U2AF35
DDX46
RNAPII
G
Log2FoldChange(IP/Input)
H
Read-through Density
Log2 Ratio Readthrough vs. Gene
Read-in Density
Log2 Ratio Read-in vs. Gene
WT-R1
WT-R2
SF3B1K700E-R1
SF3B1K700E-R2

FIGURE 4

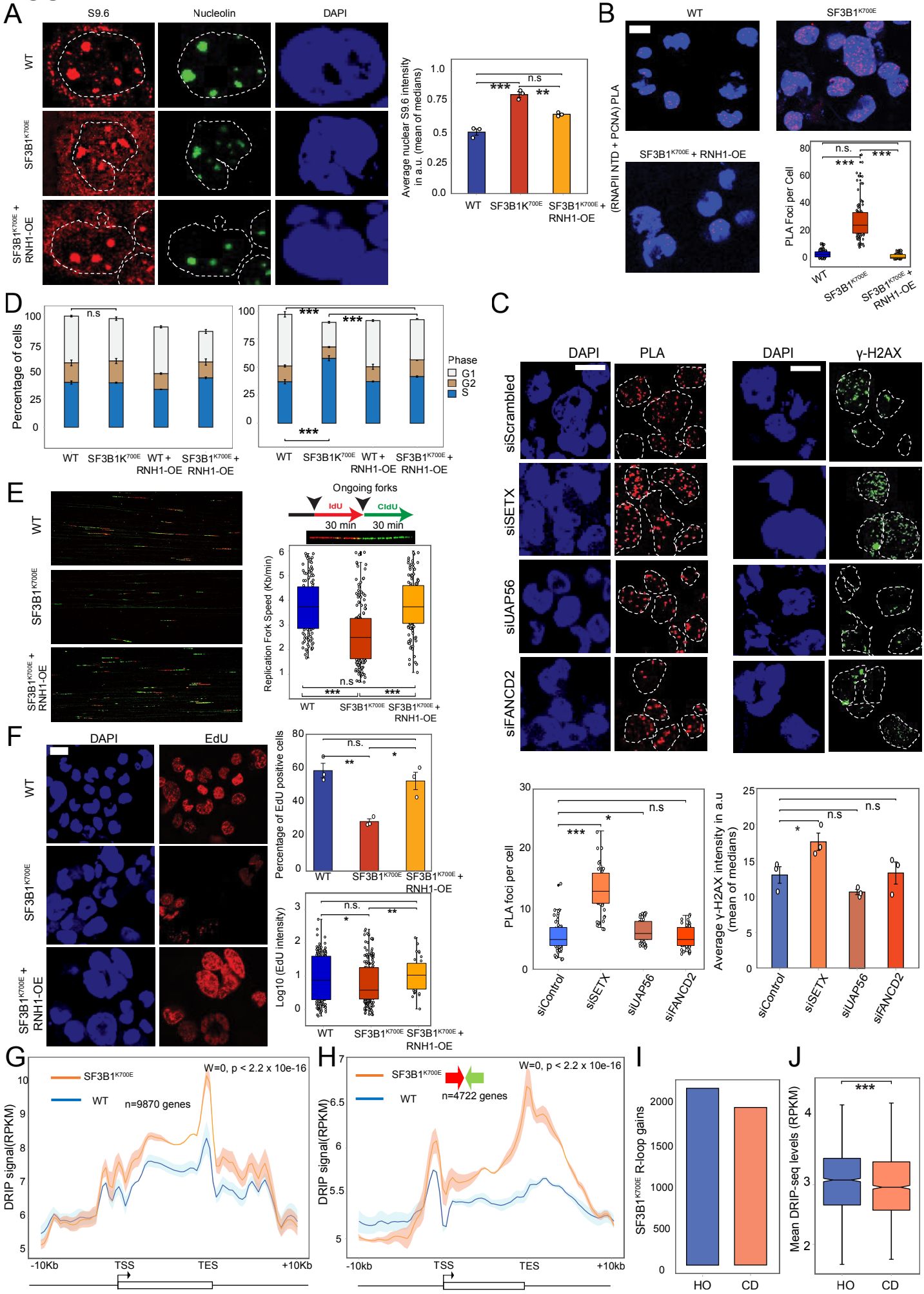
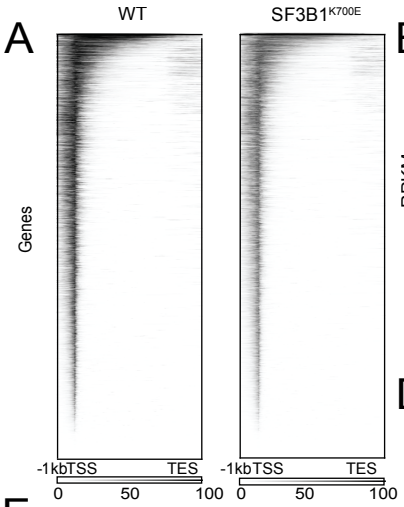
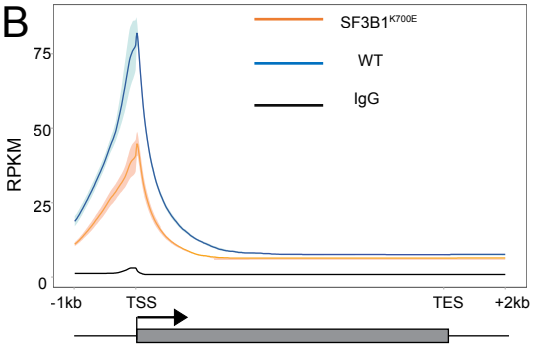
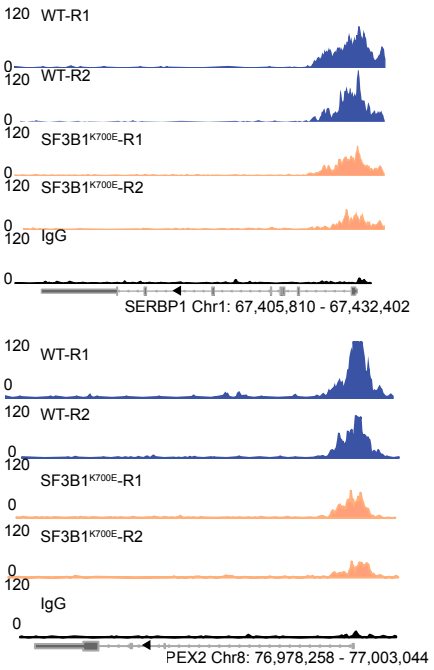
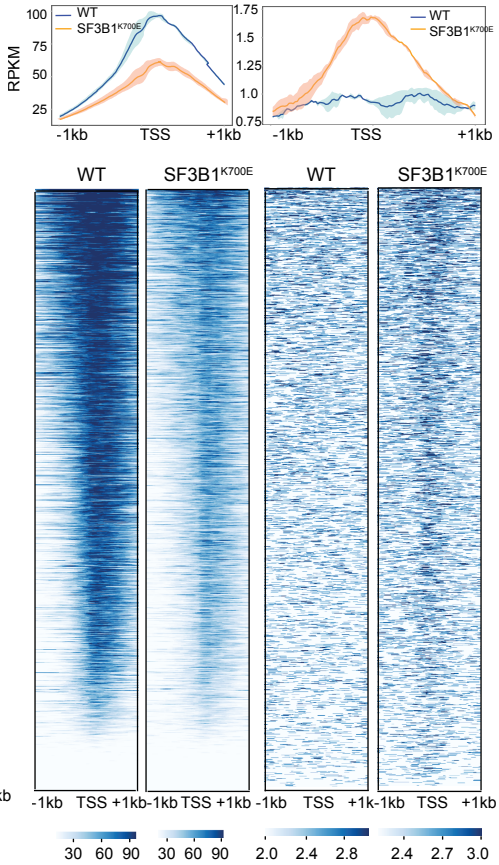

A
S9.6
Nucleolin
DAPI
WT
SF3B1K700E
SF3B1K700E + RNH1-OE
Average nuclear S9.6 intensity in a.u. (mean of medians)
n.s
***
**
B
WT
SF3B1K700E
SF3B1K700E + RNH1-OE
(RNAPII NTD + PCNA) PLA
PLA Foci per Cell
n.s.
***
C
DAPI
PLA
γ-H2AX
siScrambled
siSETX
siUAP56
siFANCD2
PLA foci per cell
Average γ-H2AX intensity in a.u (mean of medians)
siControl
D
Percentage of cells
Phase
G1
G2
S
E
Ongoing forks
IdU
CldU
30 min
Replication Fork Speed (Kb/min)
F
EdU
Percentage of EdU positive cells
Log10 (EdU intensity)
G
SF3B1K700E
WT
n=9870 genes
W=0, p < 2.2 x 10e-16
DRIP signal(RPKM)
-10Kb
TSS
TES
+10Kb
H
n=4722 genes
-10kb
+10kb
I
SF3B1K700E R-loop gains
HO
CD
J
Mean DRIP-seq levels (RPKM)

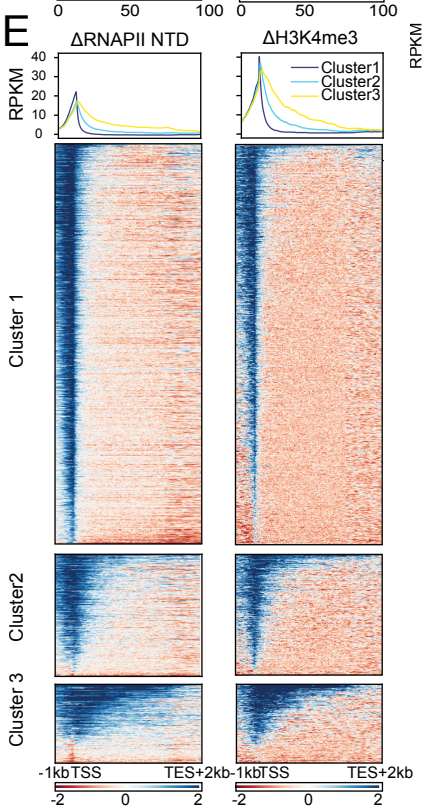
FIGURE 5

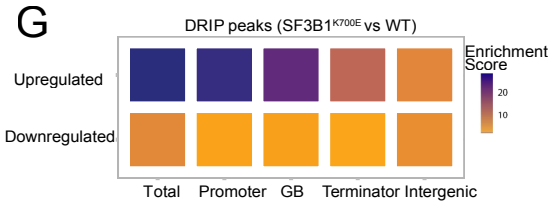

A
WT
SF3B1K700E
Genes
-1kb TSS TES
0 50 100

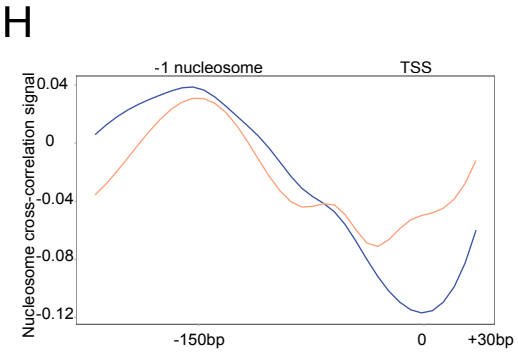

B
SF3B1K700E
WT
IgG
RPKM
75
50
25
0
-1kb TSS TES +2kb

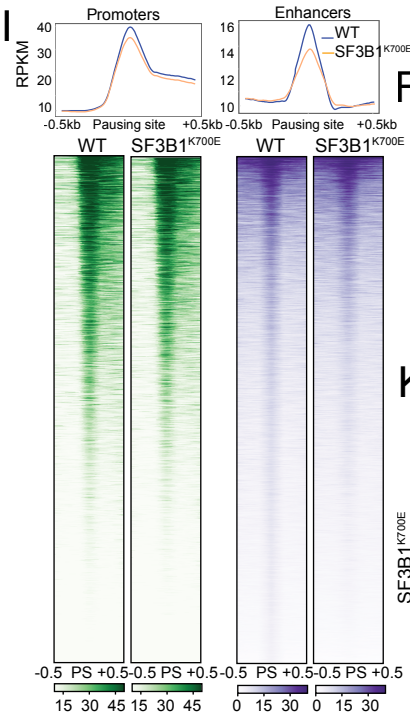

C
WT-R1
WT-R2
SF3B1K700E-R1
SF3B1K700E-R2
IgG
120
0
SERBP1 Chr1: 67,405,810 - 67,432,402
PEX2 Chr8: 76,978,258 - 77,003,044

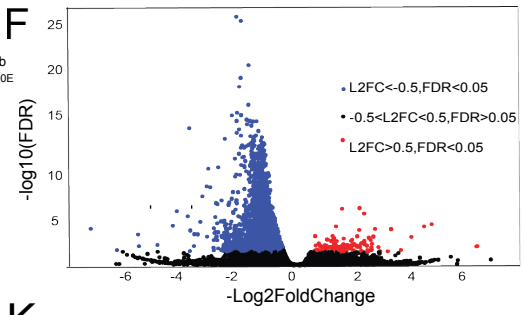

D
WT
SF3B1K700E
RPKM
-1kb TSS +1kb
30 60 90
2.0 2.4 2.8
2.4 2.7 3.0

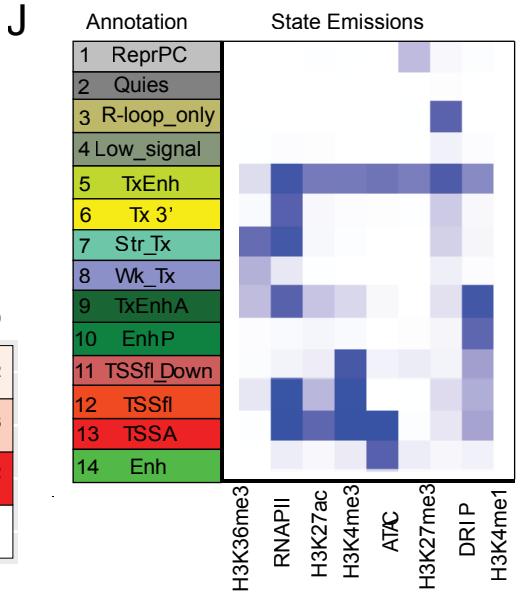

E
ΔRNAPII NTD
ΔH3K4me3
RPKM
Cluster1
Cluster2
Cluster3
Cluster 1
Cluster2
Cluster 3
-1kb TSS TES+2kb
-2 0 2


G
DRIP peaks (SF3B1K700E vs WT)
Upregulated
Downregulated
Enrichment Score
Total
Promoter
GB
Terminator
Intergenic


H
-1 nucleosome
TSS
Nucleosome cross-correlation signal
0.04
0
-0.04
-0.08
-0.12
-150bp
0
+30bp


I
Promoters
Enhancers
RPKM
WT
SF3B1K700E
-0.5kb Pausing site +0.5kb
-0.5 PS +0.5
15 30 45
0 15 30


F
L2FC<-0.5,FDR<0.05
-0.5<L2FC<0.5,FDR>0.05
L2FC>0.5,FDR<0.05
-log10(FDR)
-Log2FoldChange


J
Annotation
State Emissions
1 ReprPC
2 Quies
3 R-loop_only
4 Low_signal
5 TxEnh
6 Tx 3'
7 Str_Tx
8 Wk_Tx
9 TxEnhA
10 EnhP
11 TSSfl_Down
12 TSSfl
13 TSSA
14 Enh
H3K36me3
RNAPII
H3K27ac
H3K4me3
ATAC
H3K27me3
DRIP
H3K4me1

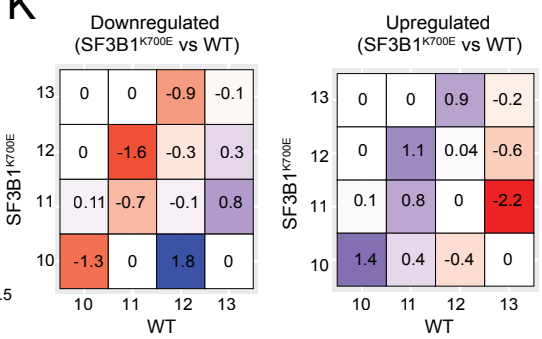
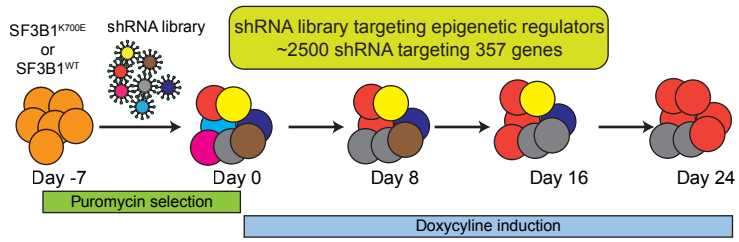
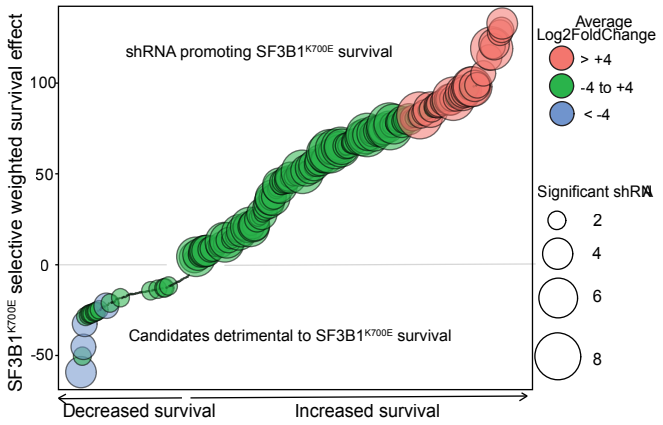

K
Downregulated (SF3B1K700E vs WT)
Upregulated (SF3B1K700E vs WT)
SF3B1K700E
WT
13 0 0 -0.9 -0.1
12 0 -1.6 -0.3 0.3
11 0.11 -0.7 -0.1 0.8
10 -1.3 0 1.8 0
13 0 0 0.9 -0.2
12 0 1.1 0.04 -0.6
11 0.1 0.8 0 -2.2
10 1.4 0.4 -0.4 0
10 11 12 13

FIGURE 6

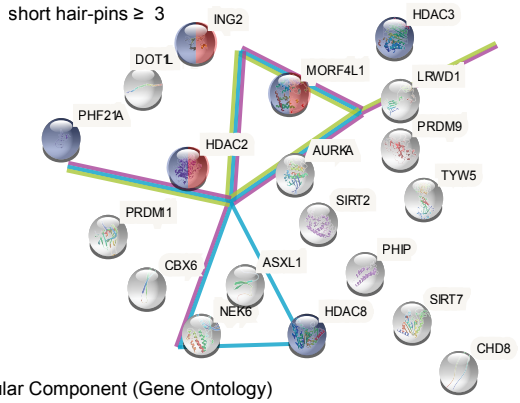
A



B



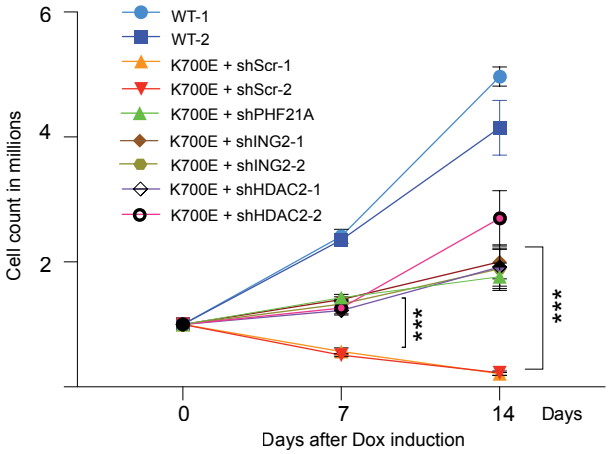
C



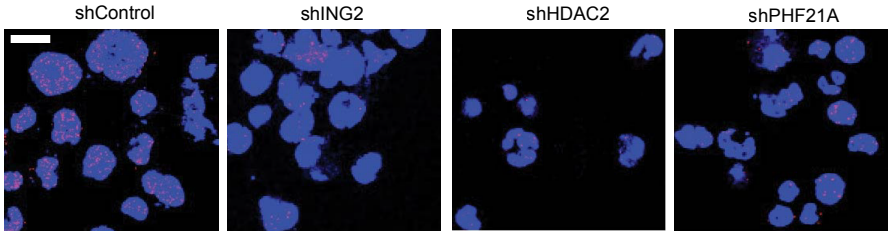
Cellular Component (Gene Ontology)

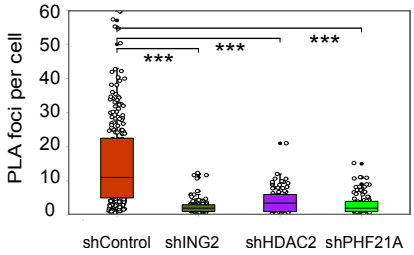
GO term	Description	False Discovery Rate
GO:0000018	Histone deacetylase complex	8.12e-08
GO:0016580	Sin3 complex	9.51e-05
GO:0031519	PcG protein complex	0.0013
GO:0000792	Heterochromatin	0.0071
GO:0072687	Meiotic spindle	0.0078

D

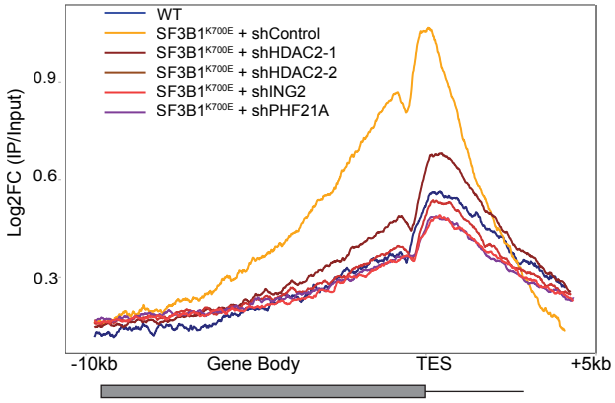


E

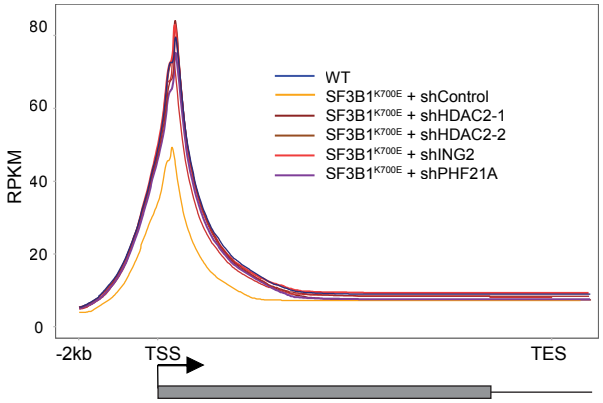




F



G



H

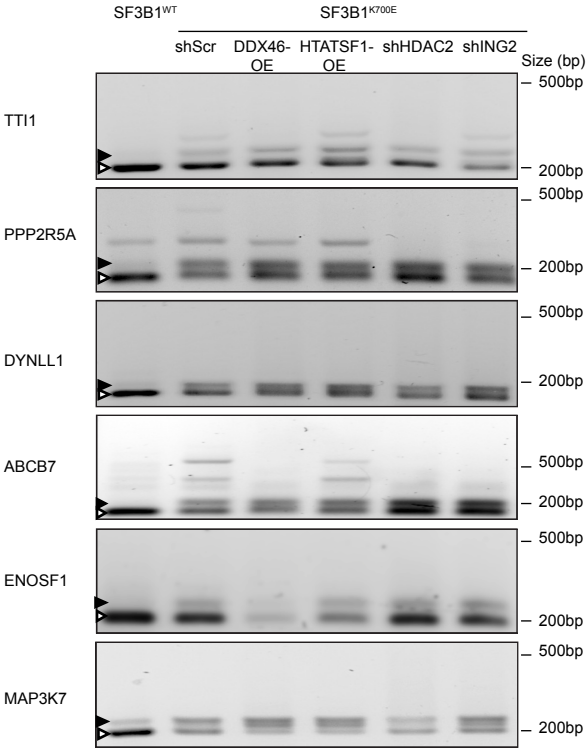
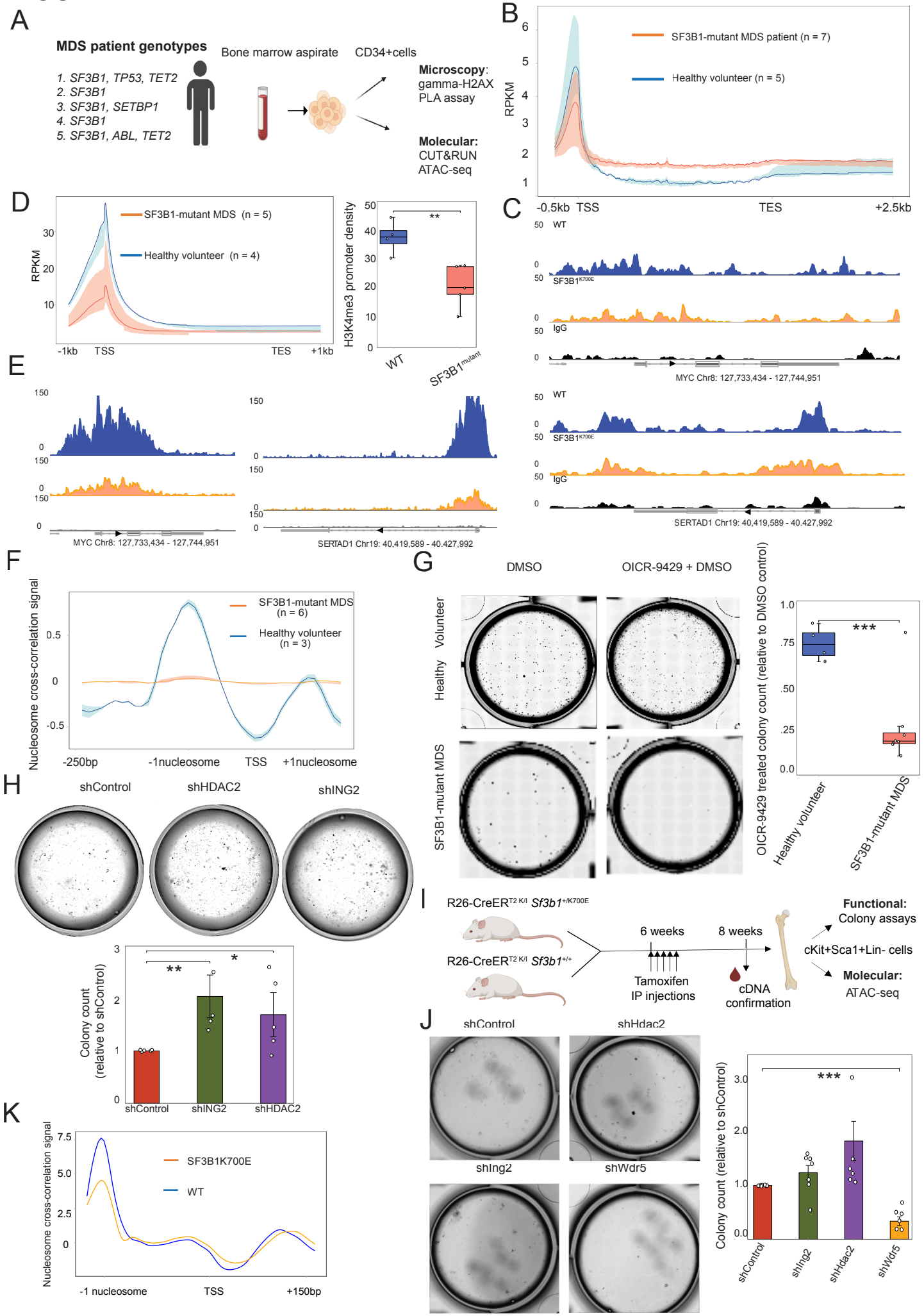


FIGURE 7


A
MDS patient genotypes
1. SF3B1, TP53, TET2
2. SF3B1
3. SF3B1, SETBP1
4. SF3B1
5. SF3B1, ABL, TET2
Bone marrow aspirate
CD34+cells
Microscopy:
gamma-H2AX
PLA assay
Molecular:
CUT&RUN
ATAC-seq
B
SF3B1-mutant MDS patient (n = 7)
Healthy volunteer (n = 5)
RPKM
-0.5kb TSS
TES
+2.5kb
C
WT
SF3B1K700E
IgG
MYC Chr8: 127,733,434 - 127,744,951
SERTAD1 Chr19: 40,419,589 - 40.427,992
D
SF3B1-mutant MDS (n = 5)
Healthy volunteer (n = 4)
RPKM
-1kb TSS
TES +1kb
H3K4me3 promoter density
WT
SF3B1mutant
E
MYC Chr8: 127,733,434 - 127,744,951
SERTAD1 Chr19: 40,419,589 - 40.427,992
F
SF3B1-mutant MDS (n = 6)
Healthy volunteer (n = 3)
Nucleosome cross-correlation signal
-250bp
-1nucleosome
TSS
+1nucleosome
G
DMSO
OICR-9429 + DMSO
Healthy Volunteer
SF3B1-mutant MDS
OICR-9429 treated colony count (relative to DMSO control)
Healthy volunteer
SF3B1-mutant MDS
H
shControl
shHDAC2
shING2
Colony count (relative to shControl)
shControl
shING2
shHDAC2
I
R26-CreERT2 K/I Sf3b1+/K700E
R26-CreERT2 K/I Sf3b1+/+
6 weeks
8 weeks
Tamoxifen IP injections
cDNA confirmation
Functional:
Colony assays
cKit+Sca1+Lin- cells
Molecular:
ATAC-seq
J
shControl
shHdac2
shIng2
shWdr5
Colony count (relative to shControl)
K
SF3B1K700E
WT
Nucleosome cross-correlation signal
-1 nucleosome
TSS
+150bp